QUANTUM MEMORY UNITS FOR COMPACT LEARNING-MEMORY SEPARATION IN NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Memory is central to sequence learning, but most neural architectures store historical information in hidden states, attention caches, or state-space transitions that are tightly coupled with feature learning. This coupling may limit the flexibility of memory storage, especially when compact long-range representations are required. Motivated by the functional separation between learning and memory in biological systems, we propose Quantum Memory Units (QMUs), a quantum-inspired memory module for hybrid classical-quantum neural networks. In our framework, a classical neural network first extracts task-relevant features, and a parameterized quantum circuit then encodes these features into compact memory states through interactions between input qubits and memory qubits. Based on this principle, we design two complementary QMUs: an episodic QMU for temporal-history compression and a semantic QMU for feature-level distribution aggregation. A Kalman-filter-based fusion mechanism is further used to combine current classical observations with quantum-inspired memory states. We evaluate QMU-enhanced networks on five memory-related tasks, including function approximation, Mackey-Glass prediction, permuted sequential MNIST, character-trajectory classification, and copy-memory. Experiments show that QMUs achieve strong performance on selected episodic memory tasks and competitive early-stage learning behavior under compact parameterization, while ROC-based analyses further indicate meaningful discriminative ability on semantic classification tasks.

1 INTRODUCTION

Learning and memory constitute fundamental cognitive functions shared by intelligent systems. Research on synaptic plasticity, axon firing, and systems-level reconfiguration has advanced our understanding of their neurobiological mechanisms and has motivated computational models of memory encoding, consolidation, and retrieval. The resulting three generations of potential-based neuron models and network architectures McCulloch & Pitts (1943); Rosenblatt (1958); Hodgkin & Huxley (1952); Maass (1997); Izhikevich (2003); Stein (1967) have been instantiated as deep neural networks (DNN) and constitute the algorithmic engine of the contemporary artificial-intelligence renaissance. In particular, recurrent neural networks (RNNs) Zaremba et al. (2014); Hochreiter & Schmidhuber (1997); Beck et al. (2024), transformers Vaswani et al. (2017), and state-space models Gu & Dao (2023) with hidden states or attentions as memory units, excel at capturing long-range, non-linear temporal dependencies. Their strength derives from synergistic memory cells and learning gating mechanisms that selectively retain, update, and blend information across time.

However, human memories are formed fundamentally differently compared to these memory mechanisms in current DNNs. Extensive neuroscience research confirms that learning and memory are not singular, homogeneous processes within the brain, but rather involve distinct separations in function, mechanisms, and neuroanatomical structures, at least exhibiting partial structural separation Kandel et al. (2014); Gazzaniga et al. (2006); Spieler et al. (2023). On the other hand, deep neural networks with state-space or attention mechanisms as their memory functions yields memory capacities too limited to accurately capture the entirety of information conveyed within a sequence, resulting in significant information loss when encoding inputs into dense vectors. In contrast, the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

brain projects information that needs to be stored to another brain mechanism at a lower resolution without retaining the original sensory input, and achieves a vast memory capacity.

Quantum systems provide high-dimensional state representations through superposition and entanglement, which motivate us to explore quantum-inspired memory modules for compact sequence representation. In this work, we do not claim practical quantum speedup or hardware-level energy advantage; instead, we focus on the representational potential of quantum-inspired memory in neural sequence learning Lyu et al. (2024a). Unlike studies that ask whether quantum phenomena occur in the brain or augment its computational repertoire Sihare (2023); Koch & Hepp (2006); Fisher (2017); Kerskens & Pérez (2022); Khrennikov & Asano (2020), we use the distinction between neural signaling and memory storage only as a computational analogy.

In this paper, we propose a quantum memory unit (QMU) and a scalable classical–quantum neural network with separable learning and memory. In the hybrid networks, classical modules extract features, while parameterized quantum circuits maintain memory states. A lightweight Kalman-filter mechanism fuses current features with memory-derived priors Kalman (1960).

Firstly, we exploit a variational quantum circuit (VQC) to engineer a QMU that encodes memory through quantum entanglement between an input (perception) qubit and memory qubits. A deep feature component from classical modules as input is prepared as a control qubit to adjust memory qubits. Regarding the two primary types of declarative memory: the episodic and the semantic, we designed two kinds of QMUs with distinct quantum circuits. The episodic QMU employs online compression Voelker et al. (2019); Gu et al. (2020) to compress rapidly arriving information via entanglement of one input qubit with multiple memory qubits, whereas the semantic QMU stores slowly varying priors through entanglement of one input qubit with a single memory qubit whose state follows a learnable probability distribution. Inspired by the functional separation between neural signaling and memory storage, we design QMU as an abstract computational memory module. The terms "input qubit" and "memory qubit" are used as a biologically inspired analogy rather than a mechanistic model of neural tissue. Secondly, for multidimensional features learned from classical deep network modules, allocating one QMU per feature component yields a scalable classical-quantum memory architecture that emulates unified yet distinct event and fact memories. Finally, we evaluate the framework on five memory-related tasks. QMU obtains its clearest gains on episodic sequence modeling, while its semantic-classification and copy-memory results are competitive or task-dependent.

The main contributions of this paper are summarized as follows:

- We introduce episodic and semantic QMU circuits that encode sequence history in compact quantum-inspired memory states separated from classical feature learning.
- We construct scalable hybrid architectures in which feature-wise QMUs interact with classical networks through a lightweight Kalman fusion mechanism.
- We evaluate QMU on five memory-related benchmarks and report both early-window performance and complete convergence curves, identifying strong episodic results as well as task-dependent semantic and copy-memory behavior.

We begin by analyzing the neural substrate for learning and memory, quantum storage capacity, and causal sequences. We then show how to build the QMU, implement its quantum circuit, and achieve appropriate behavioral functions. Subsequently, we applied the QMU to construct a hybrid classical-quantum classification neural network, thereby validating its effectiveness.

2 Related Work

2.1 Learning-memory separation in cognitive systems

Learning and memory enable intelligent systems to encode, retain, and reconstruct information. In cognitive neuroscience, memory is commonly categorized by duration and content, with long-term memory including declarative systems such as episodic and semantic memory, as well as non-declarative systems such as procedural memory Kandel et al. (2014); Gazzaniga et al. (2006). These studies suggest that learning and memory are not fully homogeneous processes, but involve partially distinct functions, mechanisms, and neural substrates. Unlike von Neumann architectures, cortical

memory is massively parallel and does not rely on discrete fetch-store cycles Koch & Hepp (2006). Dissociable amnesias further suggest that events, facts, and skills can be supported by partially separable neural systems Gazzaniga et al. (2006). Distributed neural codes also indicate that memory retrieval is not simply a process of accessing isolated containers, but depends on live, distributed representations Huffman & Guan (2025); Lyu et al. (2024b); Zhang (2024).

This functional separation motivates computational architectures in which feature learning and memory storage are not necessarily implemented by the same state variables. In contrast, most deep sequence models integrate representation learning and memory storage within recurrent hidden states, attention caches, or state-space transitions. Although such integrated designs have achieved strong performance, they may limit the flexibility of memory storage when long temporal histories need to be compressed into compact representations. Our work follows the learning-memory separation perspective by introducing a dedicated quantum-inspired memory module that operates on features extracted by classical neural networks.

2.2 Quantum-inspired compact memory representations

Quantum neural networks and quantum-inspired models have been explored as compact representation mechanisms for learning systems Lyu et al. (2024a); Li et al. (2023). An n-qubit quantum state lies in a 2^n-dimensional Hilbert space, providing a high-dimensional substrate for representing information through superposition and entanglement. However, this Hilbert-space dimension should not be interpreted as directly readable classical storage, since practical access to quantum states depends on measurement, circuit design, and sampling cost. Therefore, quantum and quantum-inspired neural models are more appropriately viewed as compact representation mechanisms rather than direct implementations of exponentially readable memory.

In this work, we use parameterized quantum circuits as quantum-inspired memory modules rather than claiming hardware-level quantum speedup or energy advantage. Specifically, the proposed QMU encodes feature components into compact memory states through interactions between input qubits and memory qubits. This design aims to investigate whether quantum-inspired state representations can provide an alternative form of memory encoding for neural sequence learning.

2.3 Sequence memory and temporal dependency modeling

Sequence modeling has been widely studied through recurrent neural networks, reservoir computing, structured memory units, transformers, and state-space models. RNNs and LSTMs store historical information in recurrent hidden states Zaremba et al. (2014); Hochreiter & Schmidhuber (1997); Beck et al. (2024), while transformers retrieve contextual information through attention mechanisms Vaswani et al. (2017). More recent state-space and structured memory models, such as Mamba, HiPPO, and LMU, aim to improve long-range dependency modeling through efficient temporal representations Gu & Dao (2023); Voelker et al. (2019); Gu et al. (2020). In particular, HiPPO and LMU formulate sequence memory as online compression, where incoming temporal signals are projected into compact coefficient representations Voelker et al. (2019); Gu et al. (2020). These methods provide important classical references for evaluating compact memory representations.

Kalman filtering provides a classical mechanism for combining predictions with observations under uncertainty Kalman (1960). In our framework, it is used as a lightweight fusion module between classical feature observations and QMU-derived memory states. Different from existing classical memory models, QMU stores feature-wise temporal or statistical information in quantum-inspired memory states separated from classical feature extraction. The episodic QMU focuses on sequence-history compression, while the semantic QMU focuses on feature-level distribution aggregation. This distinction allows us to evaluate whether separated learning-memory architectures can provide compact and task-adaptive memory representations.

3 Methods

We introduce Quantum Memory Units (QMUs) as compact memory modules that operate on feature sequences extracted by classical neural networks. Given an input sequence, a classical feature extractor first produces a sequence of feature vectors. Each feature dimension is then assigned to an

independent QMU, where a parameterized quantum circuit encodes episodic sequence reconstruction, time-series prediction, or semantic classification into memory qubits. The resulting memory states are used either for episodic sequence reconstruction or for semantic classification through a hybrid classical–quantum architecture.

3.1 FEATURE-TO-QUBIT ENCODING

Let $\mathbf{s}^{(t)} = (s_1^{(t)}, \ldots, s_d^{(t)})^\top \in \mathbb{R}^d$ denote the feature vector extracted by a classical network at time step $t \in \{0, \ldots, T-1\}$, where d is the feature dimension. For each feature component $s_i^{(t)}$, we construct a dedicated QMU. Before quantum encoding, the feature value is normalized to $\widetilde{s}_i^{(t)} \in [0, 1]$ by a task-dependent normalization function and then mapped to a rotation angle:

$$\theta_i^{(t)} = \frac{\pi}{2} \widetilde{s}_i^{(t)}. \tag{1}$$

The corresponding input qubit is prepared as a phase-free single-qubit state:

$$|\phi_i^{(t)}\rangle = \cos\left(\theta_i^{(t)}\right) |0\rangle + \sin\left(\theta_i^{(t)}\right) |1\rangle, \tag{2}$$

$$p_{i,1}^{(t)} = \sin^2\left(\theta_i^{(t)}\right), \tag{3}$$

where $p_{i,1}^{(t)}$ is the probability of observing the excited state $|1\rangle$. This encoding separates the classical feature value $s_i^{(t)}$, its normalized form $\widetilde{s}_i^{(t)}$, and the quantum rotation angle $\theta_i^{(t)}$.

3.2 QUANTUM MEMORY UNITS

A QMU is implemented as a parameterized quantum circuit consisting of one input qubit and one or more memory qubits. The input qubit receives the encoded feature component from Eq. 2, while the memory qubits maintain compact states that summarize information across time or samples. After circuit evolution, expectation values measured on the memory qubits are used as the QMU output.

Formally, for the i-th feature component, the QMU applies a parameterized circuit $\mathcal{U}_\psi$ to the input qubit and memory qubits:

$$\mathbf{m}_i^{(t)} = \mathcal{M}\left(\mathcal{U}_\psi\left(|\phi_i^{(t)}\rangle \otimes |\mathbf{h}_i^{(t-1)}\rangle\right)\right), \tag{4}$$

where $|\mathbf{h}_i^{(t-1)}\rangle$ denotes the previous memory state, $\mathcal{U}_\psi$ is a learnable quantum circuit, and $\mathcal{M}(\cdot)$ denotes measurement of memory-qubit expectation values. The measured vector $\mathbf{m}_i^{(t)}$ serves as the compact memory representation for the i-th feature component.

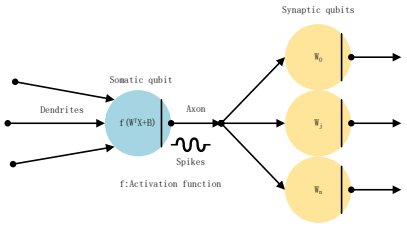
According to the type of information to be stored, we instantiate two types of QMU: episodic QMU and semantic QMU.

3.3 EPISODIC QMU FOR SEQUENCE-HISTORY COMPRESSION

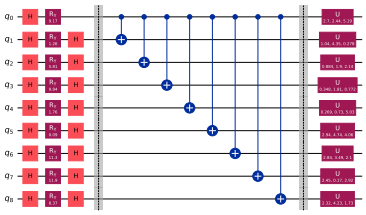
The episodic QMU is designed to compress temporal histories. For each feature component, it uses one input qubit and k memory qubits. The input qubit modulates the memory qubits through controlled interactions, allowing the circuit to accumulate information from a feature sequence over time.

For temporally structured sequences, each memory qubit interacts with the input qubit through controlled gates, as shown in Fig. 1(b). For sequences with less structured or permuted order, additional entangling gates are introduced among memory qubits to increase circuit expressiveness, as shown in Fig. 1(c). The episodic QMU evolves in a 2^k-dimensional Hilbert space, which serves as a high-dimensional representational substrate for compact sequence compression. This dimension should be understood as the state space of the circuit rather than as directly readable classical storage.

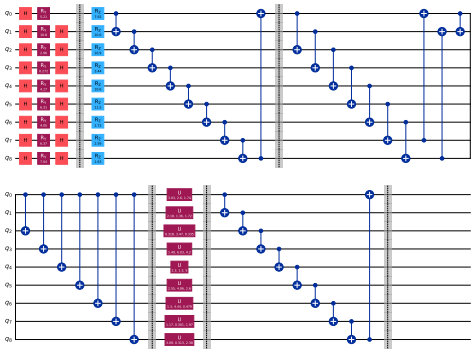
In practice, the memory-qubit expectation values are passed to a projection head or prediction head, depending on the downstream task. Thus, the episodic QMU acts as a compact temporal memory module for function approximation and time-series prediction.



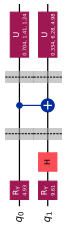
(a) QMU abstraction



(b) Structured episodic circuit



(c) Permuted episodic circuit



(d) Semantic circuit

Figure 1: Quantum memory units and their circuits. Each QMU receives one encoded input qubit and updates one or more memory qubits. The episodic QMU encodes temporal-history information through controlled interactions, while the semantic QMU accumulates feature-level statistical information through repeated input-memory interactions.

3.4 Semantic QMU for Distribution Aggregation

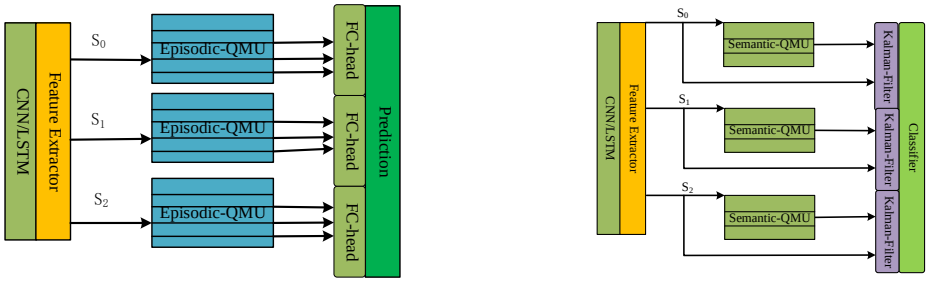
The semantic QMU is designed to aggregate feature-level statistical regularities. Unlike the episodic QMU, which uses multiple memory qubits to compress temporal history, the semantic QMU uses one input qubit and one memory qubit for each feature component, as shown in Fig. 1(d). The input qubit repeatedly interacts with the memory qubit across time or samples, allowing the memory state to encode a compact summary of feature-level statistical regularities.

In the semantic circuit, an $R_y(\theta)$ gate encodes the normalized feature value, and a universal single-qubit gate $U(\alpha, \beta, \gamma)$ provides learnable rotations. The measured expectation value of the memory qubit is used as a semantic memory representation. By assigning one semantic QMU to each feature dimension, the model obtains a feature-wise memory representation that can be used for classification.

3.5 Hybrid Classical–Quantum Networks

Based on the proposed QMUs, we construct two hybrid architectures: an episodic hybrid network and a semantic hybrid network.

Episodic hybrid network. The episodic hybrid network, shown in Figure 2(a), couples a classical feature extractor with episodic QMUs and a prediction head. The classical module, implemented by CNN or LSTM components depending on the task, first extracts feature sequences. Feature-wise episodic QMUs then compress the temporal history of each feature component into compact memory states. The resulting QMU outputs are fed into fully connected heads for sequence reconstruction or time-series prediction.



(a) Episodic hybrid network **(b)** Semantic hybrid network

Figure 2: Hybrid classical–quantum architectures. Classical CNN/LSTM modules extract feature sequences; feature-wise QMUs encode episodic or semantic memory states; prediction heads or Kalman-fusion classifiers produce the final outputs.

Semantic hybrid network. The semantic hybrid network, shown in Fig. 2(b), combines classical feature extractors, semantic QMUs, a Kalman fusion module, and a multilayer perceptron classifier. The classical network provides current feature observations, while the semantic QMU provides memory-derived priors that summarize feature-level statistical regularities. The Kalman fusion module combines these two sources of information before classification.

Kalman-filter-based fusion. To fuse current classical observations with memory-derived QMU states, we use a lightweight Kalman-filter-based fusion mechanism. In this formulation, the QMU memory state is treated as a predictive prior, while the learned classical feature is treated as a measurement. The role of this module is to stabilize the interaction between current observations and memory-derived priors, rather than to serve as an independent contribution.

Let $\mathbf{m}_k^-$ denote the memory-derived prior from QMU, $\mathbf{c}_k$ denote the current classical feature observation, $\mathbf{P}_k^-$ denote the covariance of the memory prior, $\mathbf{R}_k$ denote the observation covariance, and $\mathbf{H}_k$ denote the observation matrix. The Kalman gain and fusion update are computed as

$$\mathbf{G}_k = \mathbf{P}_k^- \mathbf{H}_k^\top \left(\mathbf{H}_k \mathbf{P}_k^- \mathbf{H}_k^\top + \mathbf{R}_k\right)^{-1}, \tag{5}$$

$$\mathbf{m}_k^+ = \mathbf{m}_k^- + \mathbf{G}_k \left(\mathbf{c}_k - \mathbf{H}_k \mathbf{m}_k^-\right), \tag{6}$$

$$\mathbf{P}_k^+ = \left(\mathbf{I} - \mathbf{G}_k \mathbf{H}_k\right) \mathbf{P}_k^-. \tag{7}$$

The fused representation $\mathbf{m}_k^+$ is then passed to the MLP classifier for semantic memory tasks.

4 Experiments

4.1 Experimental setup

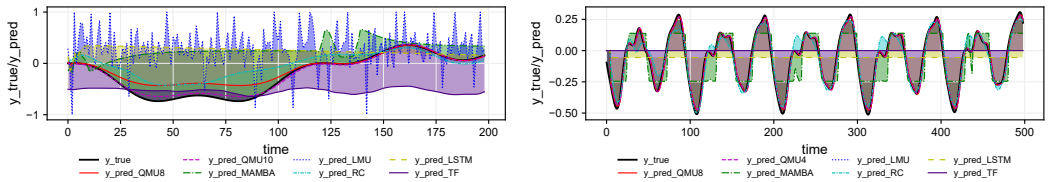
We evaluate QMU-enhanced hybrid networks on five memory-related tasks: function approximation, Mackey–Glass time-series prediction Mackey & Glass (1977), permuted sequential MNIST (PSMNIST) Goodfellow et al. (2013), character-trajectory classification Williams (2006), and copy-memory Arjovsky et al. (2016). These tasks cover continuous temporal reconstruction, chaotic time-series prediction, long sequential classification, trajectory classification, and symbolic retention.

We compare QMU-enhanced models with five classical sequence models: LSTM Hochreiter & Schmidhuber (1997), reservoir computing (RC) Jaeger (2001), LMU Voelker et al. (2019), Mamba Gu & Dao (2023), and Transformer (TF) Vaswani et al. (2017). As a scale-matched reference, classical baselines use a hidden-state dimension of 256, while QMU uses 8 memory qubits whose state lies in a 2^8-dimensional Hilbert space. We emphasize that this Hilbert-space dimension is not strictly equivalent to 256 directly readable classical hidden units; the comparison is intended to match representational scale rather than accessible classical capacity.

Episodic tasks are evaluated by mean squared error (MSE), while semantic and hybrid tasks are evaluated by classification accuracy. Since the primary goal is to evaluate compact memory and

Table 1: Early-window average performance on the five memory tasks. Lower is better for MSE; higher is better for accuracy. Copy-1 is evaluated at one interval, whereas Copy-2 averages multiple intervals.

Model	Function MSE	MG MSE	PSMNIST Acc.	Trajectory Acc.	Copy-1 Acc.	Copy-2 Acc.
LSTM	0.390	0.047	0.31	0.88	0.25	0.38
RC	0.066	0.004	0.21	0.88	0.25	0.37
LMU	0.525	0.033	0.11	0.86	0.36	0.37
Mamba	0.341	0.015	0.15	0.82	0.27	0.36
TF	0.233	0.051	**0.82**	**0.95**	**0.46**	**0.39**
QMU8	0.025	**0.001**	0.55	0.75	0.25	0.37
QMU10	**0.002**	–	–	–	–	–
QMU4	–	**0.001**	–	–	–	–



(a) Function approximation **(b)** Mackey–Glass prediction

Figure 3: Episodic-memory experiments. Curves compare QMU variants with Mamba, LMU, reservoir computing, LSTM, and transformer baselines.

rapid adaptation, Tab. 1 reports average performance over task-specific early training windows, together with convergence curves in Figures 3–5. This metric measures how quickly a model acquires useful memory representations under a limited training budget and should not be interpreted as a fully optimized final-performance benchmark.

The proposed models are implemented in PyTorch Paszke et al. (2019) using the DeepQuantum backend Zhang et al. (2025). We use tanh or softmax activations depending on the task and train all models with the Adam optimizer using its default hyper-parameters until convergence or for at most 50 epochs. Detailed task protocols are provided in Appendix A.

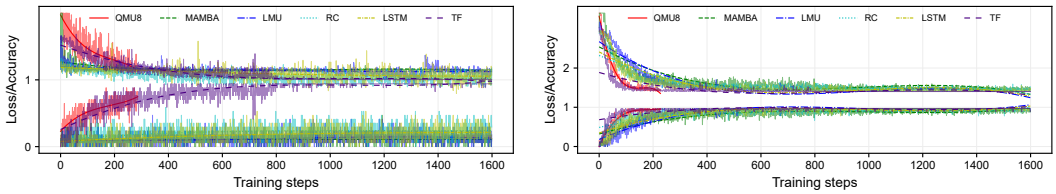
4.2 Main results

Episodic tasks. Fig. 3 and Tab. 1 show that QMU achieves its strongest results on episodic-memory tasks. On function approximation, QMU10 obtains the lowest average MSE of 0.002, followed by QMU8 with 0.025 and RC with 0.066. This result suggests that the episodic QMU is particularly suitable for compactly encoding smooth temporal patterns, where the memory state can act as a compressed representation of the underlying waveform. The improvement from QMU8 to QMU10 further indicates that increasing the number of memory qubits can provide a finer representational substrate for deterministic temporal reconstruction.

On Mackey–Glass prediction, QMU4 and QMU8 both reach an average normalized MSE of 0.001, outperforming the best classical baseline RC with 0.004. Interestingly, increasing the number of memory qubits from 4 to 8 does not further reduce the average error. This differs from the function-approximation task and suggests that performance is not determined solely by the Hilbert-space dimension. For chaotic or less regular sequences, circuit connectivity and the interaction pattern among memory qubits may be as important as the number of available memory qubits.

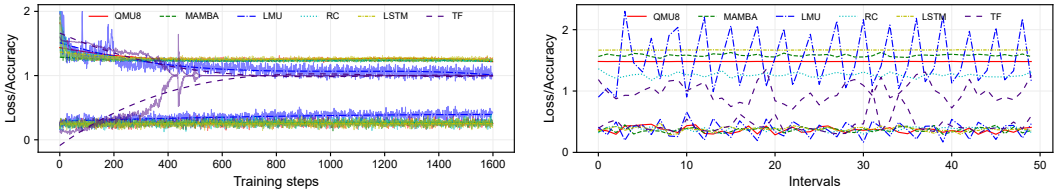
Semantic tasks. On PSMNIST, TF obtains the highest early-window accuracy of 0.82, while QMU8 reaches 0.55 and outperforms the other recurrent or state-space baselines under the deliberately short training schedule. This result indicates that the semantic QMU can extract useful discriminative information from a long permuted input sequence with a compact memory representation. However, the gap between QMU and TF also shows that attention-based models remain stronger when global token-level interactions are important for classification.

On character-trajectory classification, QMU8 obtains 0.75 average accuracy during the early window, below TF, LSTM/RC, and LMU. This task involves multi-class trajectory patterns with contin-



(a) PSMNIST **(b)** Character trajectories

Figure 4: Semantic-memory experiments. Each panel reports training loss and test accuracy for QMU8 and five classical baselines.



(a) Fixed retention interval **(b)** Interval sweep

Figure 5: Copy-memory experiments. The left panel reports training loss and copy accuracy at a fixed interval; the right panel evaluates robustness as the retention interval increases.

uous coordinate dynamics, where class-discriminative information may depend on richer temporal shape details. The relatively lower QMU accuracy suggests that the current semantic QMU, which aggregates feature-level statistical regularities, may lose part of the fine-grained temporal structure needed for this task. Therefore, semantic QMU provides a compact representation, but its discriminative capacity remains task-dependent.

Copy-memory task. On copy-memory, QMU8 obtains 0.25 accuracy at a fixed retention interval, matching LSTM and RC but trailing LMU and TF. Across multiple intervals, QMU8 reaches 0.37, equal to RC and LMU and close to LSTM and TF. This task requires the model to retain exact discrete symbols over long intervals, which is different from reconstructing continuous temporal trends or aggregating feature-level statistics. The results suggest that QMU can preserve long-range memory behavior, but its current expectation-value readout may be less suitable for exact symbol-level retrieval than architectures explicitly designed for sequence copying or attention-based recall.

4.3 Model size and compactness

Tab. 2 reports the hidden size and number of trainable parameters for each task. Classical baselines use a hidden-state dimension of 256 as a scale-matched reference, whereas QMU models use substantially fewer trainable parameters. For example, on PSMNIST, QMU8 uses 58k trainable parameters, compared with 6,950k for TF and 14,410k for LMU. On character-trajectory classification, QMU8 uses 61k trainable parameters, compared with 6,991k for TF and 14,502k for LMU. This large parameter gap indicates that QMU provides a compact memory representation rather than a heavily over-parameterized classifier.

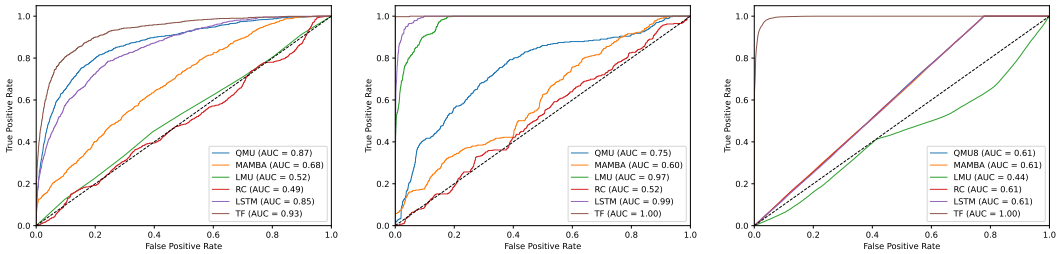
However, the results also show that compactness and best absolute accuracy should be distinguished. On episodic tasks, compact QMU models achieve the best early-window errors, suggesting that the proposed memory representation is well aligned with continuous temporal-history compression. On semantic classification and copy-memory tasks, QMU remains much smaller but does not consistently outperform the strongest baselines. Thus, the main empirical advantage of QMU in the current setting is not universal accuracy dominance, but a favorable trade-off between compact parameterization, early-stage learning, and temporal-memory representation.

4.4 Additional ROC-AUC analysis

To further examine classification behavior beyond early-window accuracy, we provide auxiliary one-vs-rest macro-average ROC-AUC results in Figure 6. On PSMNIST, QMU achieves an AUC of 0.87, which is below TF with 0.93 but higher than LSTM, Mamba, LMU, and RC. This result is con-

Table 2: Hidden size and trainable parameters in thousands for each task, reported as hidden size/parameters.

Model	Function	Mackey–Glass	PSMNIST	Trajectory	Copy memory
LSTM	256/792	256/792	256/208	256/212	256/803
RC	256/132	256/132	256/1,225	256/1,266	256/1,226
LMU	256/13,316	256/13,316	256/14,410	256/14,502	256/14,410
Mamba	256/6	256/6	256/4,246	256/38,026	256/16,965
TF	256/5,790	256/5,790	256/6,950	256/6,991	256/6,950
QMU8	8/21	8/34	16/58	16/61	16/233
QMU10	10/21	–	–	–	–
QMU4	–	4/17	–	–	–



(a) PSMNIST **(b)** Character trajectories **(c)** Copy-memory

Figure 6: Auxiliary ROC-AUC analysis on classification-related tasks. We report one-vs-rest macro-average ROC curves for PSMNIST, character trajectories, and copy-memory. The results indicate that QMU provides meaningful discriminative representations, but its advantage varies across tasks.

sistent with the accuracy comparison: QMU does not surpass the strongest transformer baseline, but its representation remains discriminative under compact parameterization. On character-trajectory classification, QMU obtains an AUC of 0.75, outperforming Mamba and RC but remaining below LMU, LSTM, and TF. This supports the observation that semantic QMU captures useful class information, but does not preserve all discriminative temporal details required by this trajectory task. On the copy-memory classification setting, QMU8 obtains an AUC of 0.61, comparable to Mamba, RC, and LSTM, but clearly below TF. This again suggests that QMU is not yet optimal for tasks requiring exact symbolic retention and retrieval.

Overall, the ROC-AUC results provide a complementary view to early-window accuracy. They show that QMU representations are not merely fitting the training curves but can provide meaningful discriminative signals. At the same time, the task-dependent AUC patterns confirm that QMU should be interpreted as a compact memory representation rather than a universally stronger classifier.

5 Conclusion

We proposed Quantum Memory Units (QMUs), a quantum-inspired memory module for compact learning-memory separation in neural networks. In the proposed hybrid classical–quantum framework, classical neural networks extract task-relevant features, while parameterized quantum circuits encode feature-wise memory states through interactions between input qubits and memory qubits. We instantiated two complementary forms of QMU: an episodic QMU for temporal-history compression and a semantic QMU for feature-level distribution aggregation. The Kalman-filter-based fusion mechanism was further used to combine classical feature observations with memory-derived QMU states. Experiments on five memory-related tasks demonstrate that QMU provides a compact and effective memory representation, especially for episodic sequence modeling. The results also show that its advantage is task-dependent on semantic classification and copy-memory tasks, where stronger transformer or task-specific memory baselines may still achieve higher absolute accuracy. These findings suggest that quantum-inspired memory units provide a promising exploratory direction for compact learning-memory separation in neural sequence models.

AI USE STATEMENT

Generative AI tools were used only for language polishing, grammar correction, mathematical notation normalization, LaTeX formatting, and figure-placement checks. They were not used to generate experimental data, compute numerical results, design the method, or draw scientific conclusions. The authors reviewed all AI-assisted edits and take full responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

We provide the experimental setup and evaluation protocol in Section 4. The five evaluated tasks are function approximation, Mackey–Glass prediction, PSMNIST classification, character-trajectory classification, and copy-memory. For each task, we specify the evaluation metric, including MSE for episodic prediction tasks and classification accuracy for semantic and hybrid tasks. The baseline models include LSTM, reservoir computing, LMU, Mamba, and transformer, with classical baselines using a hidden-state dimension of 256 as a scale-matched reference.

Detailed task protocols are reported in Appendix A, including the function-approximation target, Mackey–Glass delay setting, PSMNIST split and batch size, character-trajectory task description, and copy-memory interval settings. Implementation details are provided in Appendix B; the models are implemented in PyTorch using the DeepQuantum backend, trained with Adam for at most 50 epochs, and evaluated on CPU and GPU hardware. Task-wise hidden sizes and trainable-parameter counts are reported in Table 2. We also describe the early-window evaluation protocol used in Table 1, as well as the simulation-cost caveat and the scope of the quantum-circuit implementation in Appendix C and Appendix D. All reported QMU results are obtained through quantum-circuit simulation rather than hardware-level quantum execution. The bibliography identifies the public datasets, baseline methods, and software frameworks used in the experiments.

REFERENCES

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *International conference on machine learning*, pp. 1120–1128. PMLR, 2016.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3417.

Matthew PA Fisher. Are we quantum computers, or merely clever robots. *International Journal of Modern Physics B*, 31(7):1743001, 2017.

Michael S Gazzaniga, Richard B Ivry, and GR Mangun. Cognitive neuroscience. the biology of the mind.(2014), 2006.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. Hippo: Recurrent memory with optimal polynomial projections. *Advances in neural information processing systems*, 33: 1474–1487, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500, 1952.

Derek J Huffman and Ruijia Guan. Computational models suggest that human memory judgments exhibit interference due to the use of overlapping representations. *Psychological Review*, 2025.

Eugene M Izhikevich. Simple model of spiking neurons. *IEEE Transactions on neural networks*, 14(6):1569–1572, 2003.

Herbert Jaeger. The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German national research center for information technology gmd technical report*, 148(34):13, 2001.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960. doi: 10.1115/1.3662552.

Eric R Kandel, Yadin Dudai, and Mark R Mayford. The molecular and systems biology of memory. *Cell*, 157(1):163–186, 2014.

Christian Matthias Kerskens and David López Pérez. Experimental indications of non-classical brain functions. *Journal of Physics Communications*, 6(10):105001, 2022.

Andrei Khrennikov and Masanari Asano. A quantum-like model of information processing in the brain. *Applied Sciences*, 10(2):707, 2020.

Christof Koch and Klaus Hepp. Quantum mechanics in the brain. *Nature*, 440(7084):611–611, 2006.

Yanan Li, Zhimin Wang, Rongbing Han, Shangshang Shi, Jiaxin Li, Ruimin Shang, Haiyong Zheng, Guoqiang Zhong, and Yongjian Gu. Quantum recurrent neural networks for sequential learning. *Neural Networks*, 166:148–161, 2023.

Lei Lyu, Chen Pang, and Jihua Wang. A quantum model of biological neurons. *Neurocomputing*, 598:128223, 2024a.

Lei Lyu, Chen Pang, and Jihua Wang. Understanding the role of pathways in a deep neural network. *Neural Networks*, 172:106095, 2024b.

Wolfgang Maass. Networks of spiking neurons: the third generation of neural network models. *Neural networks*, 10(9):1659–1671, 1997.

Michael C Mackey and Leon Glass. Oscillation and chaos in physiological control systems. *Science*, 197(4300):287–289, 1977.

Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

Shyam R Sihare. Potential of quantum computing to effectively comprehend the complexity of brain. *Applied Intelligence*, 53(22):27459–27482, 2023.

Aaron Spieler, Nasim Rahaman, Georg Martius, Bernhard Schölkopf, and Anna Levina. The expressive leaky memory neuron: an efficient and expressive phenomenological neuron model can solve long-horizon tasks. In *International Conference on Learning Representations*, 2023. URL https://api.semanticscholar.org/CorpusID:259286892.

R.B. Stein. Some models of neuronal variability. *Biophysical Journal*, 7(1):37–68, 1967. ISSN 0006-3495. doi: https://doi.org/10.1016/S0006-3495(67)86574-3. URL https://www.sciencedirect.com/science/article/pii/S0006349567865743.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Aaron Voelker, Ivana Kajić, and Chris Eliasmith. Legendre memory units: Continuous-time representation in recurrent neural networks. *Advances in neural information processing systems*, 32, 2019.

Ben Williams. Character Trajectories. UCI Machine Learning Repository, 2006. URL `dataset: https://archive.ics.uci.edu/dataset/175/character+trajectories`. DOI: https://doi.org/10.24432/C58G7V.

Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *ArXiv*, abs/1409.2329, 2014. URL `https://api.semanticscholar.org/CorpusID:17719760`.

D. Zhang, X. Liu, L. Zhang, H. Wang, et al. Deepquantum: A PyTorch-native programming framework for hybrid quantum–classical machine learning. *arXiv preprint arXiv:2502.19161*, 2025. URL `https://arxiv.org/abs/2502.19161`.

Jie Zhang. How memories are stored in the brain: the declarative memory model. *arXiv preprint arXiv:2403.17985*, 2024.

A Detailed task protocols

Function approximation. For function approximation, the target signal is

$$y = \frac{\sin(x)}{4} + \frac{\sin(x/3)}{2} + \sin(x/7), \tag{8}$$

sampled at 400 points over $x \in [0, 16\pi]$. A linear decoder reconstructs each of 200 sequences from the hidden state or memory representation. The classical models use 256 hidden units, while QMU8 and QMU10 use 8 and 10 memory qubits, respectively. All models are evaluated using mean squared error (MSE).

Mackey–Glass prediction. For Mackey–Glass prediction, we integrate the delay-differential equation in the chaotic regime with delay 17. The task is to predict either the next value or the next 15 values from the previous sequence. All models are trained and evaluated using MSE. QMU4 and QMU8 are evaluated on this task to examine the influence of memory-qubit number and circuit connectivity.

Permuted sequential MNIST. For PSMNIST, each MNIST image is flattened into a length-784 sequence and permuted by a fixed permutation. We begin with the standard 60,000/10,000 MNIST train/test split and retain 10% of each split for the reported training and validation runs. All models are trained using cross-entropy loss with batch size 32. The semantic QMU uses one memory qubit per feature component, and the feature dimension is 16. The task is evaluated by classification accuracy.

Character-trajectory classification. The character-trajectory dataset contains samples represented by two spatial coordinates and pen-tip force, spanning 20 classes. All models are trained with cross-entropy loss and evaluated by classification accuracy. The semantic QMU is used to aggregate feature-level statistical regularities, and the fused representation is passed to a classifier through the Kalman fusion mechanism.

Copy memory. The copy-memory task requires the model to retain $k = 10$ digits over a sliding interval T. We evaluate cross-entropy loss and per-digit accuracy at fixed intervals $T = 120, 220, 320, 420, 520$. We also conduct an interval sweep from 60 to 5,060 to evaluate robustness as the retention interval increases. The fixed-interval result is reported as Copy-1, while the average over multiple intervals is reported as Copy-2 in Table 1.

B Additional implementation notes

The proposed models are implemented in PyTorch Paszke et al. (2019) using the DeepQuantum backend Zhang et al. (2025). The implementations have been tested on CPUs (Intel i9-11950H) and GPUs (NVIDIA GeForce RTX 3080). We use tanh or softmax activations depending on the task and train all models with the Adam optimizer using its default hyper-parameters. Training is conducted until convergence or for at most 50 epochs.

For episodic tasks, we monitor MSE. For semantic and hybrid tasks, we monitor cross-entropy loss and classification accuracy. The early-window averages reported in the main paper emphasize the first few hundred or thousand training batches because the primary objective is to assess compact memory and rapid adaptation rather than fully optimized final accuracy Lyu et al. (2024a). The convergence curves in Figures 3–5 provide the corresponding training dynamics.

C Simulation cost

We do not compare wall-clock training time as a primary metric because simulated quantum circuits are processed on conventional hardware and are therefore much slower than classical neural networks in the current software setting. In our experiments, hybrid classical–quantum training typically takes tens of hours, whereas the classical baselines require tens of minutes. Therefore, the current study focuses on the representational behavior of QMU rather than practical quantum speedup or hardware-level efficiency.

This distinction is important because the proposed QMU is evaluated through high-fidelity circuit simulation rather than real quantum hardware. The results should therefore be interpreted as evidence for quantum-inspired memory representation, not as evidence for practical acceleration or reduced energy consumption.

D Scope and limitations

This work provides a conceptual validation of quantum-inspired memory representation using circuit simulation. We do not claim practical quantum speedup, energy efficiency, or hardware-level superiority in the current setting. Since real quantum hardware relies on finite-shot measurement, the 2^k-dimensional Hilbert-space representation should be interpreted as a representational state space rather than a directly readable classical feature vector.

The current experiments are designed to study compact memory representation and early-stage learning behavior. Therefore, the reported early-window averages should not be interpreted as fully optimized final-performance benchmarks. The results show that QMU obtains its clearest gains on episodic sequence modeling tasks, while its advantage on semantic classification and copy-memory remains task-dependent.

Future work should evaluate shot efficiency, noise robustness, circuit depth, hardware feasibility, and larger-scale sequence modeling benchmarks. It would also be useful to further study circuit connectivity, measurement strategy, and task-specific readout mechanisms, especially for semantic classification and long-range symbolic recall.