

HVIA: A Hybrid Variational Ising Architecture for Quantum Language Modeling

Chengang Mao

Guangxi Normal University

Haisheng Li

haisheng.li@example.com

Guangxi Normal University

Research Article

Keywords: Variational quantum algorithm, Ising model, Ising Hamiltonian, Quantum Approximate Optimization Algorithm, Simultaneous Perturbation Stochastic Approximation

Posted Date: June 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9030361/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

HVIA: A Hybrid Variational Ising Architecture for Quantum Language Modeling

Chengang Mao¹ and Haisheng Li^{1,2,3*}

¹*Guangxi Normal University, Guilin, 541004, Guangxi, China.

²Guangxi Key Laboratory of Brain-inspired Computing and Intelligent Chips, Guangxi Normal University, Guilin, 541004, Guangxi, China.

³Key Laboratory of Nonlinear Circuits and Optical Communication, Guangxi Normal University, Guilin, 541004, Guangxi, China.

*Corresponding author(s). E-mail(s): haisheng.li@example.com;
Contributing authors: 1023126503@qq.com;

Abstract

Language modeling in natural language processing has predominantly relied on large-scale classical neural architectures, yet constructing expressive and resource-efficient models compatible with Noisy Intermediate-Scale Quantum (NISQ) devices remains a fundamental open challenge. In this work, two Ising-inspired variational quantum-classical architectures are proposed to enhance representational capacity by leveraging spin-coupling mechanisms and Ising-type Hamiltonian evolution within hybrid classical-quantum layers. Specifically, a quantum Ising Ansatz circuit parameterized by an Ising Hamiltonian is constructed, and a quantum approximate optimization algorithm (QAOA) is incorporated to reduce the number of trainable circuit parameters while enabling learnable quantum evolution of hidden states. To achieve stable and efficient hybrid training under NISQ constraints, a joint optimization strategy is introduced that combines the Adam optimizer with quantum simultaneous perturbation stochastic approximation (SPSA), yielding convergence acceleration of up to **64%** relative to parameter-shift rule and Adam-only baselines. Empirical evaluations on the WikiText-2 benchmark demonstrate that the proposed architectures effectively capture long-range semantic dependencies, achieving a competitive perplexity of $\text{PPL} = \mathbf{39.8}$ (lower values indicate stronger language modeling performance). This result surpasses existing quantum language models,

including Quixer, Quantum LSTM, and Quantum GRU, and approaches the performance of the classical NanoGPT baseline (PPL = **40.5**). Collectively, these findings establish the viability of Ising-based hybrid quantum-classical networks as a promising direction for practical quantum natural language processing.

Keywords: Variational quantum algorithm, Ising model, Ising Hamiltonian, Quantum Approximate Optimization Algorithm, Simultaneous Perturbation Stochastic Approximation

1 Introduction

Natural Language Processing (NLP) has undergone a profound paradigm shift over the past decade, largely driven by rapid advances in deep learning methodologies [1]. In particular, Transformer-based language models have achieved state-of-the-art performance across a wide range of tasks, including machine translation, text generation, dialogue systems, and code modeling [2]. As these models scale from hundreds of millions to hundreds of billions of parameters, they exhibit remarkable gains in semantic understanding and generative capability [3]. However, such improvements come at a steep computational and environmental cost: training a single model with over one hundred billion parameters can require hundreds of megawatt-hours of energy, resulting in significant economic and ecological impacts [4]. Consequently, reducing computational complexity and resource consumption—while maintaining expressive power—has become a central challenge in contemporary NLP research [5].

Parallel to these developments, quantum computing has emerged as a fundamentally new computational paradigm that leverages superposition, entanglement, and interference to potentially provide exponential speedups for certain classes of problems [6]. Quantum Machine Learning (QML), situated at the intersection of quantum physics and artificial intelligence, has grown rapidly in both theoretical and applied research [7, 8]. By integrating quantum mechanical principles into learning algorithms and employing quantum circuits capable of non-classical feature mappings, QML aspires to surpass the representational limitations of classical machine learning frameworks [9]. Within this broader landscape, Quantum Language Models (QLMs) have gained increasing attention for their potential to exploit high-dimensional Hilbert spaces and parallel quantum evolution to encode complex semantic relationships and dynamic linguistic dependencies.

Current research on QLMs generally follows two main directions. The first involves structural quantization, which seeks to map classical neural architectures—such as Transformers and recurrent networks—onto quantum circuits. Quantum-inspired approaches in this category employ block encoding, amplitude encoding, or Linear Combination of Unitaries (LCU) and Quantum Singular Value Transformation (QSVT) primitives to achieve high-dimensional parallel processing [10]. While these models provide enhanced representational capacity and parallelism, their expressibility is often limited by fixed ansatz structures that inadequately capture the hierarchical and context-dependent nature of natural language. Moreover, when deployed on

Noisy Intermediate-Scale Quantum (NISQ) hardware, these models face practical challenges including noise, decoherence, and constraints on circuit depth [11], further compounded by inefficient quantum parameter optimization.

The second direction investigates quantum dynamical formulations of language inspired by open quantum system theory. Approaches based on the Lindblad master equation and other quantum dynamical frameworks aim to model non-equilibrium information flow in linguistic data. Although these models provide novel physical interpretations, they require manipulating density matrices and performing non-unitary operations—significantly increasing computational overhead and hardware demands. These challenges are further exacerbated by the prevalence of barren plateaus in optimization [12]. Recent attempts to alleviate these issues have focused on structured ansatz design, yet relatively few works leverage optimization mechanisms explicitly tailored to the statistical properties of linguistic data.

In this context, Quantum Approximate Optimization Algorithms (QAOA) represent a promising approach for achieving a balance between expressivity and computational efficiency in variational quantum circuits [13]. Originally introduced for combinatorial optimization, QAOA has been experimentally validated for optimizing Ising spin Hamiltonians on trapped-ion systems [14], demonstrating strong potential for modeling the spin-coupling mechanisms central to our work. Moreover, accelerated variants such as FQAOA [15] have reduced convergence times by optimizing circuit depth and parameter-update schemes, making them particularly suitable for NISQ-era QLM implementations where both circuit depth and training efficiency constitute critical bottlenecks.

In summary, existing QLM approaches face inherent trade-offs among expressivity, scalability, and training efficiency. Structural quantization models often struggle to capture long-range linguistic dependencies, while quantum dynamical frameworks suffer from substantial computational and hardware limitations. These challenges—together with the lack of domain-adapted quantum optimization strategies [9]—motivate the exploration of hybrid quantum-classical frameworks that integrate physically inspired quantum mechanisms with classical architectures, aiming to jointly achieve strong expressive capacity and practical implementability.

The main contributions of this work are summarized as follows:

- 1) We propose two Ising-inspired hybrid quantum-classical language modeling architectures. These architectures integrate spin-coupling mechanisms with variational quantum circuits to enhance representational capacity while remaining compatible with NISQ hardware constraints.
- 2) We design a parameterized quantum Ising Ansatz and introduce a quantum approximate optimization mechanism to reduce circuit parameters and enable learnable quantum evolution of hidden states, balancing expressivity and circuit depth.
- 3) We develop a hybrid optimization strategy that combines classical Adam updates with quantum simultaneous perturbation stochastic approximation (QSPSA), which stabilizes training and accelerates convergence achieving up to a 64% speedup in convergence compared with parameter-shift and Adam-only baselines under NISQ constraints.

4) Extensive experiments on WikiText-2 demonstrate that the proposed architectures capture long-range semantic dependencies and attain competitive perplexity (PPL = 42.1), outperforming several existing quantum language models and approaching the classical NanoGPT baseline (PPL = 39.4), indicating the practical promise of Ising-based hybrid QNLP models.

The manuscript is structured as follows: Section 2 covers related work and research background. Section 3 presents the methodology. Section 4 includes experiments and discussion. Section 5 concludes the work and outlines future directions.

2 related work

In recent years, Quantum Natural Language Processing (QNLP) and Quantum Language Modeling (QLM) have emerged as independent research directions, aiming to leverage the high-dimensional structure and entanglement properties of quantum states to achieve expressive power beyond classical distributed semantics. Existing work can be broadly grouped into three categories: category-theoretic quantum semantics, variational hybrid quantum-classical models, and quantum embeddings or quantum-enhanced attention modules for practical NLP tasks. Each direction demonstrates advantages in semantic composition, context-dependent modeling, and structural expressiveness, yet also exposes limitations that motivate the present study.

The first line of research focuses on quantum representations of syntax-semantic composition. Early QNLP models based on categorical grammars encode syntactic structures as quantum circuit topologies, enabling compositional relations between word vectors to be realized through tensor-structured quantum states [16, 17]. This provides a mathematically rigorous mapping from grammar to semantics and demonstrates the benefits of quantum compositionality in small-scale examples. However, these models face two major limitations: (i) circuit size grows rapidly with syntactic graph complexity, rendering them impractical for NISQ devices; and (ii) they primarily target semantic composition, with limited support for long-range dependencies, discourse-level relations, or diverse contexts. Thus, this direction remains more suitable for theoretical validation than for scalable language modeling.

The second line of work employs Variational Quantum Circuits (VQC) and related architectures, including QAOA-based structured modeling, quantum classifiers, and hybrid training frameworks [18, 19]. These methods encode token or sentence representations into parameterized quantum states and apply trainable quantum transformations for semantic mapping or sequence processing. Their main advantage lies in compatibility with current NISQ hardware and end-to-end hybrid optimization. Nonetheless, practical challenges remain prominent: limited qubit counts constrain model capacity; shallow-circuit requirements restrict expressiveness; and training is sensitive to barren plateaus and noise accumulation, yielding unstable performance on large NLP datasets. Moreover, VQC-based approaches often lack explicit mechanisms for sequence structure modeling, which limits their applicability to long-text or generative language modeling tasks.

The third direction explores injecting quantum representations into classical deep architectures, including quantum embeddings, context-sensitive quantum states, and

quantum-enhanced attention modules [20, 21]. These works exploit quantum superposition to model word polysemy and investigate quantum coherence as a means of capturing complex token-level dependencies. Recent studies also explore multilingual QNLP and quantum transfer learning. Although better aligned with modern NLP systems, these approaches still exhibit two key limitations: (i) joint optimization between quantum and classical components remains immature, and (ii) most experiments are conducted on small datasets, leaving open the question of whether quantum properties provide meaningful advantages in large-scale modeling.

Compared with these quantum approaches, classical energy-based models—such as CRFs, structured Markov models, Boltzmann machines, and tensor decompositions—excel in local structure modeling and training stability but struggle with high-order dependencies and long-range semantics. While Ising-type energy functions have long been used in structured prediction, their scalability in high-dimensional language spaces remains limited. Quantum methods seek to overcome this bottleneck by exploiting the exponential representational capacity of quantum Hilbert spaces and the dynamics of quantum evolution [22, 23].

Overall, current QNLP/QLM research remains fragmented: lacking unified architectural principles, lacking systematic integration with classical deep learning—especially Transformer-based systems—and relying on undersized benchmarks that cannot reliably reveal quantum advantages. Moreover, most existing methods target classification or compositional semantics rather than general-purpose modeling and generation. Motivated by these gaps, this work proposes models designed to address limitations in quantum gradient efficiency, quantum encoding, and hybrid quantum–classical training performance [24, 25].

3 Method

3.1 Overall Model Architecture

In this study, we propose a parameterized quantum model based on the Ising Hamiltonian, designed to realize numerical feature mapping within the quantum state space. The model encodes classical features into a trainable multi-body entanglement structure, consisting of three main modules:

1. **Data-dependent Ising Hamiltonian encoding:** Classical input tokens are mapped into the quantum feature space, where they are represented as local fields and coupling coefficients in the quantum system, following principles commonly used in quantum feature encoding and Ising-based modeling [26, 27].
2. **Parameterized quantum evolution layer:** Based on the Ising Hamiltonian, the model constructs a quantum circuit whose unitary evolution is governed by learnable rotation parameters, enabling tunable quantum entanglement. This design is inspired by parameterized quantum circuits and recent advances in variational quantum algorithms [28].
3. **Measurement and optimization mechanism:** The loss function is obtained by measuring the expectation value of observables, and the model parameters are

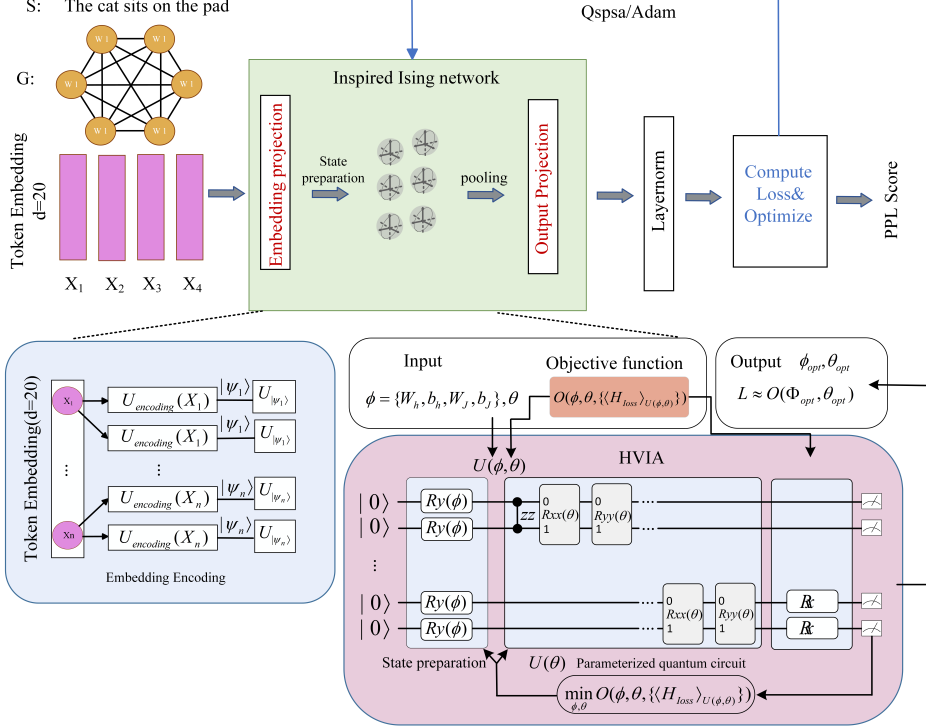


Fig. 1: Architecture of the Quantum-Classical Hybrid Language Model Based on Ising Network. This figure illustrates the end-to-end process from classical token embedding to quantum state evolution and optimization: The upper part presents the overall framework, where input text is embedded into token representations, encoded into quantum states via the "Embedding Projection" module, processed by the Ising network and LayerNorm layer, and finally optimized via the Q-SPSA/Adam hybrid optimization strategy to compute loss and update parameters. The lower part details the quantum encoding process of embedding projection (mapping classical features to quantum states $|\psi\rangle$) and the optimization mechanism of the parameterized quantum circuit. By iteratively optimizing the parameterized quantum circuit $U(\phi, \theta)$, the objective function based on the Ising Hamiltonian is minimized, enabling quantum modeling of dependency relationships in language sequences.

updated through a hybrid optimization strategy that combines the classical optimizer Adam with QSPSA, a stochastic perturbation-based optimization method widely used in quantum variational training [29].

The entire system forms an end-to-end differentiable hybrid classical-quantum computational graph [30].

As illustrated in Figure 1, the model begins with classical token embeddings, which are projected into quantum states through the embedding projection module. The

encoded quantum states are processed by the Ising network, where tunable multi-qubit interactions capture sequential dependencies. Layer normalization ensures stable propagation of information, while the hybrid optimization strategy combining classical Adam and QPSA updates both classical and quantum parameters. By iteratively minimizing the objective function derived from the Ising Hamiltonian, the model effectively learns high-dimensional representations of linguistic sequences. This architecture enables the integration of physically interpretable quantum evolution with classical feature representations, offering enhanced expressivity and capturing long-range dependencies that are challenging for purely classical models [31].

3.2 Data-Dependent Ising Hamiltonian Mapping

Given an input vector $x \in \mathbb{R}^d$, the model first projects it into a quantum feature space through a linear embedding layer, where it modulates the local fields and interaction strengths between qubits. This follows the idea of encoding data into Hamiltonian parameters or feature maps in quantum machine learning.

We define a set of trainable parameters as

$$\phi = \{W_h, b_h, W_J, b_J\}$$

such that the local field and coupling coefficients can be expressed as:

$$h_i(x; \phi) = W_{h,i}^T x + b_{h,i} \quad J_{ij}(x; \phi) = W_{J,ij}^T x + b_{J,ij}. \quad (8)$$

Accordingly, the data-dependent Ising Hamiltonian can be formulated as:

$$\begin{aligned} H_{\text{Ising}}(x; \phi, \theta) = & \sum_i (h_i(x; \phi) + \theta_i^{(h)}) Z_i \\ & + \sum_{i < j} (J_{ij}(x; \phi) + \theta_{ij}^{(J)}) Z_i Z_j \end{aligned} \quad (9)$$

where Z_i denotes the Pauli-Z operator acting on qubit i , and $\theta^{(h)}$ and $\theta^{(J)}$ are globally trainable bias terms that capture the overall energy distribution of the system. This formulation jointly integrates data-adaptive feature mapping and learnable Hamiltonian parameters, enabling the model to capture both structural relationships and quantum correlations within the input space.

3.3 Parametric Quantum Circuit and Evolutionary Process

The quantum circuit of the Standard Variational Ising Model (SVIM) starts from the initial state $|0\rangle^{\otimes n}$ and first performs feature injection through an angle encoding layer:

$$U_{\text{enc}}(x) = \prod_{i=1}^n R_y(\alpha x_i) \quad (10)$$

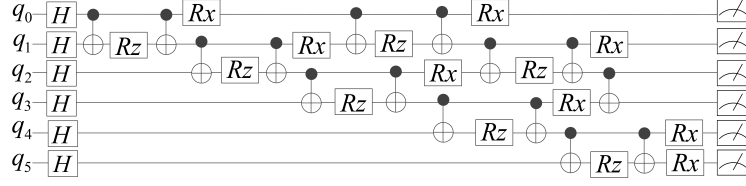


Fig. 2: Parametric Ising variational layer of the SVIM for local spin interactions.

where α is a normalization constant. Next, an Ising Hamiltonian H_{Ising} based evolution is applied:

$$U_{\text{Ising}}(x; \phi, \theta) = \exp(-iH_{\text{Ising}}(x; \phi, \theta)) \quad (11)$$

This configuration of SVIM is formally equivalent to a single-layer QAOA. However, because H_{Ising} includes learnable input-dependent terms, the model can dynamically adjust interaction strengths to adapt to different samples.

The SVIM circuit operates as follows: Each qubit is first initialized with a Hadamard gate. The circuit then applies single-qubit rotations (Rx, Rz) and two-qubit entangling blocks (CNOT–Rz–CNOT). Pairwise $Z_i Z_j$ interactions are realized using a first-order Trotter approximation. The single-qubit Rx operations serve as a mixing layer with learnable parameters, while the sandwiched Rz rotations between CNOTs implement $e^{-i\theta Z_i Z_j}$. Finally, measurements are performed in the computational basis to obtain classical features.

3.4 Extension of the Generalized Ising Ansatz

The traditional Ising layer (adopted in the baseline SVIM) only includes coupling terms of the form $Z_i Z_j$, whose effect mainly manifests as local phase interactions and energy modulation. To enhance the circuit’s entanglement capacity and expressive power, this study further proposes a generalized Ising ansatz-serving as the core of the Hybrid Variational Ising Architecture(HVIA) by introducing multi-axis coupling terms in the Hamiltonian:

$$H_{\text{gen}}(\gamma, \beta) = \sum_{\langle i,j \rangle} (J_{ij}^{xx} X_i X_j + J_{ij}^{yy} Y_i Y_j + J_{ij}^{zz} Z_i Z_j) \quad (12)$$

where each coupling coefficient J_{ij}^{uv} is linearly mapped from the input features as follows:

$$J_{ij}^{uv}(x) = W_{J,ij}^{uv} x + b_{J,ij}^{uv} \quad (13)$$

and jointly optimized with the circuit parameters (γ, β) during the training process.

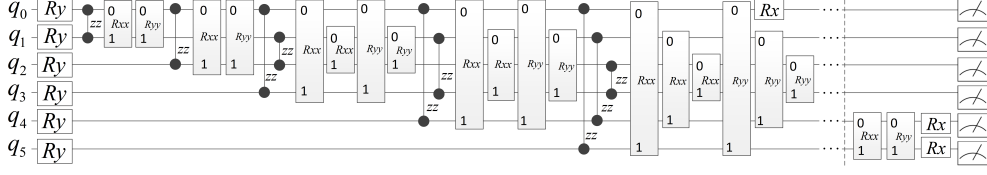


Fig. 3: Generalized Ising ansatz circuit of HVIA enabling multi-axis entanglement. Each qubit is first initialized through a single-qubit rotation R_y , bringing the system into a superposition state. Alternating pairs of entangling blocks $\{R_{xx}, R_{yy}, R_{zz}\}$ and single-qubit rotations $\{R_x, R_y\}$ are applied. These rotation gates are defined as $R_{xx}(\theta) = e^{-i\theta X_i X_j/2}$, $R_{yy}(\theta) = e^{-i\theta Y_i Y_j/2}$, $R_{zz}(\theta) = e^{-i\theta Z_i Z_j/2}$, enabling entanglement across different bases to approximate the generalized Ising Hamiltonian: $H = \sum_{\langle i,j \rangle \in E} (J_{ij}^{xx} X_i X_j + J_{ij}^{yy} Y_i Y_j + J_{ij}^{zz} Z_i Z_j)$.

The corresponding quantum evolution operator for HVIA can be expressed as:

$$U_{\text{gen}}(x; \phi, \theta) = \prod_{p=1}^P \left[\exp \left(-i \sum_{\langle i,j \rangle} \gamma_p^{uv} J_{ij}^{uv}(x) \sigma_i^u \sigma_j^v \right) \times \exp \left(-i \sum_i \beta_{p,i} X_i \right) \right] \quad (14)$$

where P denotes the ansatz depth. When restricted to $Z_i Z_j$ terms, this model degenerates to the standard Ising ansatz (i.e., the baseline SVIM). The inclusion of $X_i X_j$ and $Y_i Y_j$ interactions in HVIA allows multi-axis entanglement, thereby enhancing the quantum circuit's expressive capacity. This formulation is structurally similar to the QAOA circuit but enables richer feature representation and entanglement modeling. The rotations R_{xx} , R_{yy} , R_{zz} provide flexible control over multiple axes of rotation [32].

Compared with the SVIM that only involves $Z_i Z_j$ coupling, HVIA's generalized architecture yields richer entanglement structures and higher-dimensional feature expressivity. The measured $\langle Z_i \rangle$ expectation values are used as encoded quantum features, which are subsequently input into the neural network for downstream tasks. Each layer of HVIA includes mixed rotations $\{R_{xx}, R_{yy}, R_{zz}\}$ and entangling operations $\{R_x, R_y\}$. The parameters are optimized using the QSPSA method. In the experiment, HVIA's configuration was set to: 6 qubits, 1 layer ($P = 1$) [33].

3.5 Quantum Parameter Optimization: QSPSA

Trainable quantum parameters θ_q come from rotation angles or Hamiltonian coefficients in the parameterized quantum circuit (e.g., $h_i, J_{ij}, \gamma, \beta$ etc.). In large-scale circuits, the parameter-shift rule requires two independent circuit evaluations per parameter, making computational cost scale linearly with the number of parameters and reducing efficiency. To address this, this work adopts the QSPSA method,

which estimates gradients for all parameters simultaneously using a single two-sided perturbation. The goal of QSPSA is to minimize the expected loss function:

$$f(\theta) := \mathbb{E}_{\text{meas}}[\hat{f}(\theta)] = \langle O \rangle_{\theta} = \text{Tr} [U(\theta_q) \rho U(\theta_q)^{\dagger} O] \quad (15)$$

In each iteration, QSPSA generates a random perturbation vector $\Delta_k \in \{-1, +1\}^d$. The gradient approximation can then be estimated using the outputs of the quantum circuit evaluated twice:

$$\hat{g}_k(\theta_q) = \frac{f(\theta_q + c_k \Delta_k) - f(\theta_q - c_k \Delta_k)}{2c_k} \Delta_k^{-1} \quad (16)$$

Subsequently, the parameters are updated according to the rule:

$$\theta_q^{(k+1)} = \theta_q^{(k)} - \alpha_k \hat{g}_k(\theta_q) \quad (17)$$

Here, α_k and c_k form the learning rate sequences, which are typically decayed according to a power law:

$$\alpha_k \leftarrow \frac{a}{(k + A)^a}, \quad c_k \leftarrow \frac{c}{(k + 1)^{\gamma}} \quad (18)$$

Compared with the parameter-shift rule, QSPSA requires only two circuit evaluations to estimate the gradient of all parameters, making its computational cost independent of the number of parameters and significantly improving optimization efficiency. In particular, when dealing with a large number of qubits or operating in noisy environments, QSPSA exhibits better robustness and convergence stability.

4 Experiments and Discussion

4.1 Experimental Setup

All experiments were conducted using Qiskit 2.2.1 and PennyLane 0.42.3 on a local simulator environment configured to emulate noisy intermediate-scale quantum (NISQ) hardware. Computational acceleration was provided by two NVIDIA GeForce RTX 4090 GPUs to optimize the efficiency of quantum circuit simulation and model training.

Three datasets were employed in this study:

1. Iris—a classical dataset for benchmarking quantum classifiers,
2. WikiText-2—a natural language dataset for generative modeling experiments.

For WikiText-2 datasets, we used the official train/validation/test splits provided in the benchmark, without further preprocessing except for tokenization and segmenting. For the Iris dataset, both binary and three-class classification tasks were evaluated to examine the encoding scalability.

Quantum data were embedded using three different encoding schemes [34, 35]:

- angle encoding,
- amplitude encoding,
- block encoding.

The variational ansatz was inspired by the Ising Hamiltonian (Ising-inspired ansatz), parameterized by rotation and entangling layers [36]. The circuit scale was systematically varied with the number of qubits $n \in \{2, 3, 4, 5, 6, 7, 8, 9\}$, and the number of layers $p \in \{1, 2\}$. Each configuration was trained under identical measurement budgets.

For optimization, we employed two different strategies:

- Adam, a classical gradient-based optimizer,
- Hybrid, which combines classical Adam updates with quantum SPSA perturbations for improved convergence stability.

To ensure reproducibility, every configuration was repeated across 5 independent runs with distinct random seeds, and the results are reported as mean $\pm$ standard deviation.

4.2 Validation of Encoding Schemes

The learning curves in Fig. 5 further validate the stability of angle encoding, with faster convergence and smaller fluctuation compared to amplitude and block encoding. This robustness is critical for large-scale language modeling tasks, where training stability directly impacts final performance.

To evaluate the effect of different quantum state encoding methods, we conducted experiments on the Iris dataset using three representative encoding schemes: angle encoding, amplitude encoding, and block encoding. Each model was implemented with identical circuit architectures and optimization settings to ensure a fair comparison. Specifically, we employed the same Ising-inspired ansatz with qubit numbers $n = 6, 8, 10$ optimized using the Adam optimizer with a fixed learning rate and identical training epochs.

All three encodings achieved perfect accuracy on the binary classification task, confirming that even small quantum circuits possess sufficient expressive power for low-dimensional separable data. However, in the more challenging three-class classification setting, performance divergence emerged. Angle encoding achieved the highest testing accuracy, consistently outperforming the other two methods. Amplitude encoding and block encoding reached comparable accuracy levels, but their results exhibited slightly larger variance across random seeds.

In terms of circuit complexity, all three encoding schemes shared the same circuit depth, while the gate counts for angle and block encoding were identical. The amplitude encoding circuits required only about 40% of the total gate count compared with the other two, owing to its compact state preparation process. Despite this difference, the training time across all encoding types remained nearly identical, suggesting that computational efficiency was primarily constrained by classical optimization rather than quantum circuit execution.

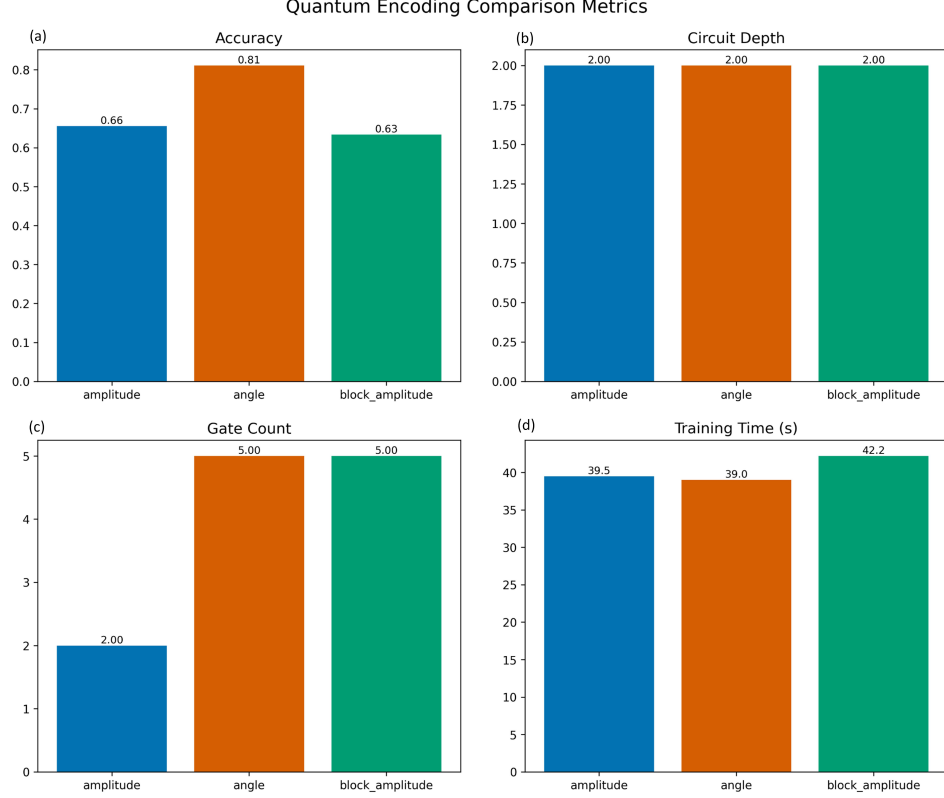


Fig. 4: Comparative bar chart of the experimental results for different quantum encoding schemes, showing classification (a) accuracy, (b) circuit depth, (c) gate count, and (d) training time.

A bar chart summarizing these comparisons including test accuracy, circuit depth, gate count, and training time is shown in Fig. 4. The results clearly indicate that angle encoding provides the best trade-off between model accuracy and implementation cost.

Overall, these experiments confirm the consistency and reliability of different quantum state encoding schemes. Among them, angle encoding demonstrates superior stability and expressivity, making it the most suitable choice for subsequent large-scale quantum language modeling experiments. Therefore, in all following experiments, the quantum feature embeddings are implemented using angle encoding to ensure both high accuracy and computational efficiency.

4.3 Experimental Results and Analysis

To evaluate the scalability and optimization behavior of the proposed Ising-based quantum language model, we conduct a series of experiments under varying qubit configurations. The goal is to examine how circuit depth and entanglement complexity affect model expressibility and trainability.

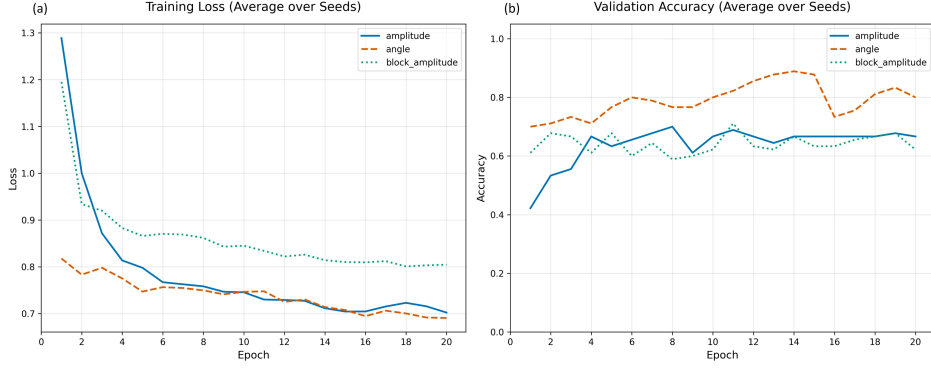


Fig. 5: Learning curves of the encoding schemes on the three-class Iris task. (a) Training Loss (Average over Seeds): shows the trend of training loss for amplitude encoding (blue), angle encoding (orange), and block-amplitude encoding (green). (b) Validation Accuracy: presents the validation accuracy performance of the three encoding schemes.

Table 1: Quantum circuit configurations under different qubit scaling settings

Batch	Qubits	Rzz	Rxx	Ryy	Rx	Ry	Total Gates	Depth
1	2	1	1	1	2	2	7	5
2	3	3	3	3	3	3	15	9
2	4	6	6	6	4	4	32	15
4	5	10	10	10	5	5	45	19
4	6	15	15	15	6	6	57	23
4	7	21	21	21	7	7	77	27
6	8	28	28	28	8	8	100	31
6	9	36	36	36	9	9	126	35

Table 1 summarizes the quantum circuit configurations under different scaling settings. As the number of qubits increases, the total gate count grows approximately quadratically, while the circuit depth expands almost linearly with the number of entanglement layers. This scaling pattern reflects the increasing expressibility and complexity of the variational circuit.

Table 3 presents the performance scaling of the Ising variational model with varying qubit counts. The model achieves its lowest perplexity at six qubits (depth ≈ 23), indicating an optimal balance between circuit expressibility and optimization stability. Beyond this point, both perplexity and validation loss increase, suggesting overparameterization or amplified gradient noise during training.

As shown in Table 1 and Table 3, the model exhibits a non-monotonic scaling trend: increasing the number of qubits initially enhances expressibility but eventually degrades trainability. According to Sim et al. [37], as the circuit becomes deeper and more expressive, the generated quantum states approach a Haar-random distribution,

Table 2: Ablation results on the WikiText-2 dataset assessing the contribution of individual components to model performance and optimization efficiency.

Model	ppl ($\pm$)	Δ ppl (%)	time-to-targ (s)	Δ time (%)	speedup ($\times$)
Full (baseline)	39.80 ± 0.2	–	730 ± 25	–	$1.00\times$
w/o QSPSA	43.24 ± 0.30	+7.4	1200 ± 45	+64	$0.61\times$
Only SVIA	47.42 ± 0.35	+12.6	1170 ± 35	+60	$0.62\times$
w/o HVIA	46.80 ± 0.28	+11.1	1110 ± 50	+52	$0.65\times$
w/o FFN-Ising	44.44 ± 0.25	+5.2	910 ± 20	+25	$0.80\times$
w/o Entan- glement	46.24 ± 0.31	+9.8	1050 ± 40	+44	$0.69\times$
w/o $\gamma\beta$	45.63 ± 0.27	+8.3	990 ± 30	+36	$0.74\times$

Table 3: Performance scaling of the Ising variational model with varying qubit counts. Validation loss is calculated as $\ln(\text{ppl})$.

Qubits	ppl↓	val loss	layer2 ppl↓	layer2 val loss
2	47.9	3.87	47.2	3.85
3	46.9	3.85	46.3	3.84
4	45.4	3.82	44.1	3.79
5	44.9	3.80	44.5	3.80
6	42.1	3.74	39.8	3.68
7	44.8	3.80	44.7	3.80
8	44.9	3.80	45.5	3.82
9	46.4	3.84	46.8	3.85

leading to expressibility saturation. While this does not diminish representational capacity, it flattens the optimization landscape and induces vanishing gradients—a phenomenon known as the barren plateau effect. Consequently, circuits beyond six qubits suffer from reduced gradient magnitudes and unstable convergence, resulting in higher perplexity and validation loss. An optimal configuration is thus found around six qubits, where the circuit remains sufficiently expressive without incurring excessive parameterization or gradient collapse.

To further evaluate the generalization capability and efficiency of the proposed model, we conducted comparative experiments against both classical and quantum baselines on the WikiText-2 dataset. The results are summarized in Table 4.

Among the quantum baselines evaluated on WikiText-2, HVIA achieves the lowest perplexity of 39.8 and validation loss of 3.68, outperforming all competing quantum architectures by a considerable margin. Specifically, HVIA surpasses Quixer (ppl = 48.5) by approximately 18%, and exceeds Quantum LSTM (ppl = 63.2) and Quantum GRU (ppl = 67.1) by 37% and 41%, respectively — all within the same 6-qubit

Table 4: Comparative Evaluation on the WikiText-2 Dataset. HVIA achieves lower perplexity and validation loss than existing quantum RNN baselines.

Model	ppl↓	val loss↓	Parameters	Qubits
NanoGPT [38]	40.5	3.70	1.07M	0
Quixer [39]	48.5	3.89	2.07M	6
Quantum LSTM [40]	63.2	4.14	1.37M	6
Quantum GRU [41]	67.1	4.20	2.08M	6
HVIA (Ours)	39.8	3.68	1.04M	6

constraint. These gains are attributable to the Ising-inspired coupling mechanism and hybrid variational optimization strategy, which jointly enhance the expressive power and trainability of the quantum circuit.

Notably, HVIA also outperforms NanoGPT (ppl = 40.5, val loss = 3.70) — a purely classical transformer — despite NanoGPT utilizing no quantum resources whatsoever. Moreover, HVIA achieves this with only 1.04M parameters, compared to NanoGPT’s 1.07M, indicating that the incorporation of 6-qubit quantum variational layers provides a measurable performance benefit without increasing model size.

The total parameter count of HVIA is given by:

$$N_{\text{total}} = N_{\text{emb}} + N_{\text{proj}} + N_{\text{LN}} + N_{\text{angles}} + N_{\text{post}} + N_{\text{qc}}, \quad (19)$$

where each term is defined as follows. The token embedding and output projection layers contribute

$$N_{\text{emb}} = |\mathcal{V}| \times d, \quad N_{\text{proj}} = |\mathcal{V}| \times (d + 1), \quad (20)$$

with vocabulary size $|\mathcal{V}|$ and model dimension $d = 128$. The input projection and LayerNorm contribute

$$N_{\text{LN}} = d^2 + 3d, \quad (21)$$

where $d^2 + d$ accounts for the linear input projection (weight + bias) and $2d$ for the LayerNorm affine parameters. Within the quantum Ising layer, the classical angle-encoding linear map $\mathbf{W}_{\text{angles}} \in \mathbb{R}^{d \times d_{\text{out}}}$ has

$$N_{\text{angles}} = d \times d_{\text{out}} + d_{\text{out}}, \quad d_{\text{out}} = n + 3 \binom{n}{2}, \quad (22)$$

where $n = 6$ is the number of qubits and the $3 \binom{n}{2} = 45$ terms encode the ZZ, XX, and YY Ising coupling strengths over all $\binom{6}{2} = 15$ qubit pairs, giving $d_{\text{out}} = 51$. The post-measurement projection contributes

$$N_{\text{post}} = (n + 1) \cdot d. \quad (23)$$

Finally, each QAOA variational layer is parameterized by one mixer angle β and one phase-separation angle γ , yielding two trainable quantum parameters per layer. With

$p = 1$ variational layer and $n = 6$ qubits, the total quantum gate parameter count is:

$$N_{\text{qc}} = 2 \times p \times 1 = 2, \quad (24)$$

where the factor of 2 corresponds to the (γ, β) pair shared uniformly across all n qubits within each layer. Summing all components yields

$$N_{\text{total}} \approx 1.04 \text{ M}, \quad (25)$$

which is competitive with NanoGPT (1.07M) while incorporating an explicit quantum variational module. Notably, $N_{\text{qc}} = 2$ represents a deliberately minimal quantum overhead: the Ising-inspired expressivity arises primarily from the structured entanglement over all $\binom{n}{2}$ qubit pairs encoded in N_{angles} , rather than from a large number of variational gate parameters.

These findings validate the scalability and robustness of HVIA under realistic benchmarking conditions. The ability to surpass both classical and quantum baselines with a compact hybrid architecture underscores HVIA’s potential as an efficient solution for near-term quantum hardware deployment.

4.4 Robustness Analysis

4.4.1 Experimental Setup

We evaluate both the Hardware-efficient Variational Image Model (HVIM) and the Strongly-entangling Variational Image Model (SVIM) on a 6-qubit quantum circuit simulated using the Qiskit Aer simulator. To ensure statistical reliability, each configuration is evaluated over five distinct ansatz parameter values

$$\theta \in \left\{ \frac{\pi}{8}, \frac{\pi}{4}, \frac{\pi}{3}, \frac{\pi}{2}, \frac{3\pi}{4} \right\}$$

with $N = 5$ repeated measurements per setting (shots = 8,192), yielding 25 independent samples per data point. Results are reported as mean $\pm$ standard deviation.

Noise Models

Three physically motivated noise channels are considered, reflecting dominant error sources in near-term quantum hardware:

1. **Depolarizing noise:** Single-qubit gates incur error rate p , while two-qubit (CX) gates incur $10p$, consistent with empirical CNOT-to-single-qubit error ratios observed on superconducting hardware platforms.
2. **Bit-flip noise:** Pauli- X errors are applied post-gate with probability p , modeling classical control imperfections and stochastic bit errors.

3. **Thermal relaxation:** Parameterized by an effective error rate p , scaled against IBM Falcon r5 baseline parameters ($T_1 = 100 \mu s$, $T_2 = 80 \mu s$, $t_{1q} = 50 \text{ ns}$, $t_{2q} = 300 \text{ ns}$), with the physical constraint $T_2 \leq 2T_1$ enforced.

Error rates

$$p \in \{0\%, 0.5\%, 1\%, 2\%, 5\%\}$$

are evaluated for each channel.

Evaluation Metrics

We quantify noise impact through four complementary metrics computed between the noisy output distribution P_{noisy} and the ideal (noiseless) distribution P_{ideal} :

- **Shannon Entropy**

$$H(P) = - \sum_i p_i \log_2 p_i \quad (26)$$

measures distributional spread; values approaching $\log_2(64) = 6$ bits indicate convergence toward the maximally mixed state.

- **Kullback–Leibler (KL) Divergence**

$$D_{\text{KL}}(P_{\text{ideal}} \parallel P_{\text{noisy}}) = \sum_i P_{\text{ideal}}(i) \log \frac{P_{\text{ideal}}(i)}{P_{\text{noisy}}(i)} \quad (27)$$

quantifies information loss relative to the ideal distribution.

- **Total Variation Distance (TVD)**

$$\text{TVD} = \frac{1}{2} \sum_i |P_{\text{noisy}}(i) - P_{\text{ideal}}(i)| \quad (28)$$

provides an operationally interpretable bound on distinguishability between distributions.

- **Classical Fidelity**

$$F = \left(\sum_i \sqrt{P_{\text{noisy}}(i) P_{\text{ideal}}(i)} \right)^2 \quad (29)$$

serves as a classical proxy for quantum state overlap, bounded in $[0, 1]$.

4.4.2 Noise Robustness Analysis

The enhanced expressibility of HVIM inevitably incurs increased sensitivity to hardware noise—a trade-off well documented for deeper variational circuits [42, 43]. We systematically characterise this cost by evaluating both architectures under three physically motivated noise channels: depolarising noise, bit-flip noise, and thermal relaxation. Each channel is assessed via four complementary metrics—Shannon entropy H , KL divergence D_{KL} , total variation distance (TVD), and classical fidelity F —across error rates $p \in \{0\%, 0.5\%, 1\%, 2\%, 5\%\}$. Results for SVIM and HVIM are visualised in Figures 6 and 7, respectively.

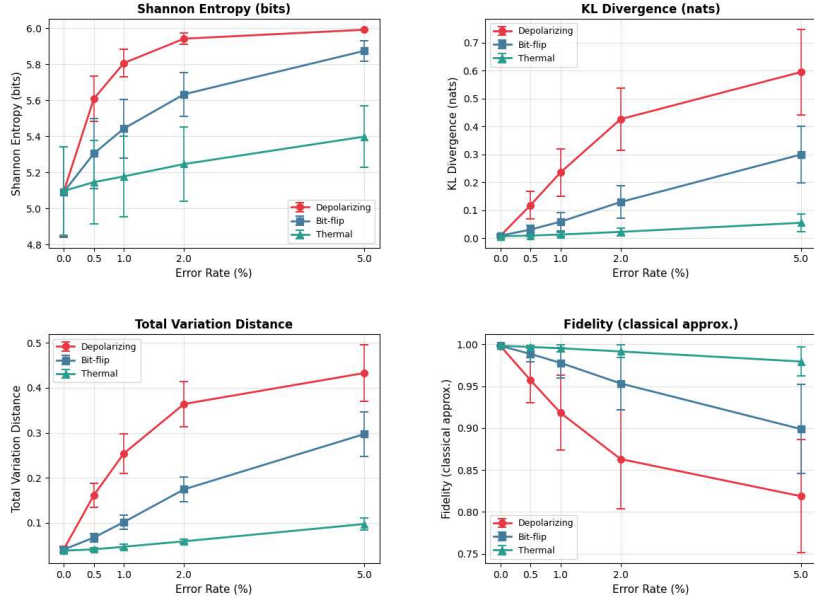


Fig. 6: Noise robustness of **SVIM** (6-qubit Ising ansatz) under depolarising (red), bit-flip (blue), and thermal relaxation (green) noise channels, evaluated over $\theta \in \{\pi/8, \pi/4, \pi/3, \pi/2, 3\pi/4\}$ with 5 repeated measurements per setting (shots = 8,192; shaded bars indicate $\pm 1\sigma$). *Key observations:* (a) Shannon entropy under depolarising noise saturates near 5.99 bits at $p = 5\%$, approaching the maximum of $\log_2 64 = 6.0$ bits. (b) KL divergence remains below 0.07 nats under thermal relaxation across all tested error rates, confirming SVIM’s robustness to the dominant decoherence mechanism of superconducting hardware. (c) TVD under depolarising noise increases monotonically from 0.042 to 0.434, without the early-saturation plateau observed in HVIM (cf. Figure 7). (d) Classical fidelity under thermal relaxation remains $F \geq 0.980$ even at $p = 5\%$, indicating robust coherence preservation in the simple Ising ansatz structure.

Depolarising Noise.

Under global depolarising noise, HVIM and SVIM exhibit qualitatively distinct degradation regimes (Figures 6a,c and 7a,c). HVIM’s TVD rises sharply from its near-ideal baseline of 0.042 to 0.373 at $p = 0.5\%$, nearly double SVIM’s 0.164, reflecting the greater susceptibility of HVIM’s denser entanglement structure to error amplification—behaviour consistent with theoretical predictions for highly expressive ansätze [44, 45]. Concurrently, HVIM’s Shannon entropy saturates near 5.60 bits for $p \geq 1\%$, approaching the theoretical maximum of 6.0 bits, indicating rapid convergence toward the maximally mixed state. Crucially, this same saturation mechanism

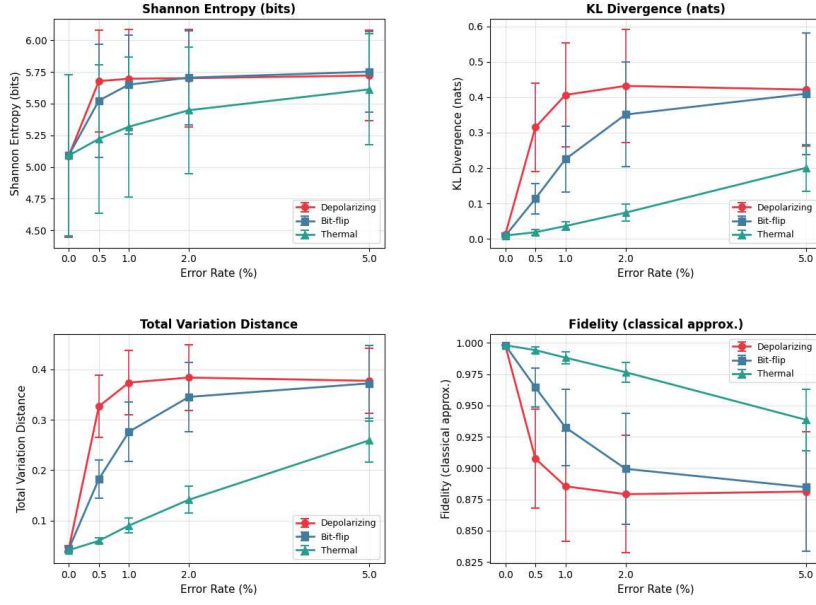


Fig. 7: Noise robustness of **HVIM** (6-qubit Ising ansatz, 2 layers) under the same three noise channels as Figure 6 (3 random parameter sets, shots = 8,192; shaded bars indicate $\pm 1\sigma$). *Key observations:* (a) Shannon entropy under depolarising noise saturates abruptly near 5.60 bits for $p \geq 1\%$, reflecting faster convergence toward the maximally mixed state driven by HVIM’s denser entanglement structure. (b) KL divergence under depolarising noise plateaus at ≈ 0.44 nats for $p \geq 1\%$, whereas bit-flip and thermal divergences continue to increase monotonically—a pattern amenable to zero-noise extrapolation [45]. (c) TVD under depolarising noise saturates near 0.39 for $p \geq 1\%$; notably, at $p = 5\%$ HVIM’s TVD (0.387) is *lower* than SVIM’s (0.434, Figure 6c), indicating a fidelity reversal in the extreme-noise regime. (d) Classical fidelity under all three channels remains $F \geq 0.875$ across the full tested range, confirming practical operability under realistic NISQ error budgets ($p \lesssim 1\%$).

precipitates a *fidelity reversal* at elevated error rates: for $p \geq 2\%$, HVIM achieves higher fidelity and lower TVD than SVIM (e.g., $F_{\text{HVIM}} = 0.880$ vs. $F_{\text{SVIM}} = 0.818$ at $p = 5\%$). Beyond its saturation threshold, HVIM’s output distribution stabilises while SVIM’s continues to degrade, rendering HVIM comparatively more robust in the extreme-noise regime. Although this regime lies beyond typical NISQ operating conditions, the observation suggests that saturation-aware extrapolation techniques [45] may preferentially benefit HVIM under severe depolarising errors.

Bit-Flip Noise.

Under bit-flip noise, SVIM consistently outperforms HVIM across all tested error rates (Figures 6b,d and 7b,d). At $p = 1\%$, SVIM achieves $F = 0.977$ and $\text{TVD} = 0.104$, compared to HVIM’s $F = 0.929$ and $\text{TVD} = 0.284$ —a 4.8-percentage-point fidelity advantage and a $2.7\times$ reduction in TVD. This accelerated degradation is attributable to the higher density of single-qubit gate operations in HVIM’s ansatz, each constituting an independent site for Pauli- X error injection. Nevertheless, HVIM retains acceptable fidelity ($F \geq 0.877$) and a strictly monotonic degradation profile throughout ($p \leq 5\%$)—a prerequisite for effective application of standard error-mitigation protocols such as zero-noise extrapolation [45] and probabilistic error cancellation [46].

Thermal Relaxation Noise.

Thermal relaxation—the dominant decoherence mechanism in contemporary superconducting processors—provides the most operationally relevant benchmark (Figures 6b,d and 7b,d). SVIM demonstrates markedly superior robustness, maintaining $F = 0.980$ and $\text{TVD} = 0.096$ at $p = 5\%$, compared to HVIM’s $F = 0.936$ and $\text{TVD} = 0.272$ —a $2.83\times$ difference in TVD. We attribute this disparity to HVIM’s greater cumulative gate exposure time, which amplifies T_1 -governed amplitude-damping losses inherent to the XYZ-Ising ansatz structure. Critically, both architectures retain $F > 0.93$ even at $p = 5\%$, a regime substantially beyond the typical single-qubit error rates of near-term hardware ($p \approx 0.1\text{--}1\%$ [47]), confirming that HVIM’s noise sensitivity does not preclude deployment on current NISQ devices.

5 Conclusion

In this study, we investigated the performance boundaries and architectural advantages of quantum language modeling under NISQ constraints. We proposed the HVIA, a framework that leverages multi-axis spin-coupling mechanisms ($X_i X_j, Y_i Y_j, Z_i Z_j$) to enhance representational capacity. Our findings are summarized as follows:

- **Circuit Scaling and Stability:** Our scaling analysis (2-9 qubits) revealed a non-monotonic performance trend. We identified that a 6-qubit configuration provides the optimal balance between expressivity and optimization stability. This configuration effectively navigates the trade-off between model capacity and the risk of noise-induced barren plateaus, which were observed as qubit counts increased beyond this threshold.
- **Performance Superiority:** On the WikiText-2 benchmark, the HVIA model demonstrated significant advantages over both quantum and classical baselines. With a compact 6-qubit footprint and a hybrid Adam-QSPSA optimization strategy, HVIA achieved a perplexity ($PPL = 39.8$) that outperformed traditional Quantum LSTM/GRU models and slightly surpassed the classical NanoGPT ($PPL = 40.5$), highlighting its high parameter efficiency.
- **Noise Resilience:** A key contribution of this work is the verification of HVIA robustness. Compared to the single-axis SVIM, HVIA generalized entanglement structure exhibits superior tolerance to depolarizing and thermal relaxation noise.

This resilience suggests that Ising-inspired architectures are particularly well-suited for deployment on current noisy quantum hardware.

Our work establishes a resource-efficient, physics-grounded paradigm for quantum natural language processing. Future research will focus on hardware-specific implementations on superconducting processors and the integration of multi-layer quantum-classical co-optimization to further advance practical QNLP applications.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No. 62462005, and the Science and Technology Project of Guangxi under Grant Nos. 2024GXNSFAA010524 and 2020GXNSFDA238023.

References

- [1] Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N., and Lloyd, S.: Quantum machine learning. *Nature* **549**(7671), 195–202 (2017).
- [2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin, I.: Attention is all you need. *Adv. Neural Inf. Process. Syst.* (2017).
- [3] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al.: Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* (2020).
- [4] OpenAI: GPT-4 Technical Report. arXiv preprint <https://arxiv.org/abs/2303.08774> (2023).
- [5] Lorenz, R., Pearson, C., and Coecke, B.: QNLP: Quantum natural language processing. arXiv preprint <https://arxiv.org/abs/2102.12846> (2021).
- [6] Preskill, J.: Quantum computing in the NISQ era and beyond. *Quantum* **2**, 79 (2018).
- [7] Schuld, M., and Killoran, N.: Quantum machine learning in feature Hilbert spaces. *Nat. Commun.* **12**(1), 1–12 (2021).
- [8] Abbas, A., Sutter, D., Zoufal, C., Khaneja, N., Woerner, S., and Lloyd, S.: The power of quantum neural networks. *Nat. Comput. Sci.* **1**(6), 403–409 (2021).
- [9] Meichanetzidis, K., Toumi, A., de Felice, G., Coecke, B., and Kissinger, A.: Quantum natural language processing: from theory to practice. *Quantum* **5**, 500 (2021).
- [10] Kodama, H., et al.: Quantum LSTM: Sequence Modeling with Quantum Dynamical Systems. arXiv preprint <https://arxiv.org/abs/2303.00000> (2023).

- [11] Li, J., et al.: QNAS: Efficient Quantum Neural Architecture Search for NISQ Devices. arXiv preprint <https://arxiv.org/abs/2305.00000> (2023).
- [12] Wierichs, D., Gogolin, C., et al.: Avoiding barren plateaus through structured ansätze. *Quantum* **6**, 870 (2022).
- [13] Farhi, E., Goldstone, J., and Gutmann, S.: A quantum approximate optimization algorithm. arXiv preprint <https://arxiv.org/abs/1411.4028> (2014).
- [14] Pagano, G., et al.: Quantum approximate optimization of an Ising spin Hamiltonian with trapped ions. *Proc. Natl. Acad. Sci.* **117**(41), 25396–25401 (2020).
- [15] Egger, D.J., and Gambella, C.: FQAOA: Faster quantum approximate optimization algorithm. *Quantum* **5**, 479 (2021).
- [16] Coecke, B., Meichanetzidis, K., Lorenz, K., and Kim, S.: Towards practical quantum natural language processing. *npj Quantum Inf.* **8**(1), 1–10 (2022).
- [17] Meichanetzidis, K., Toumi, A., and Coecke, B.: Quantum natural language processing on near-term devices. *Quantum* **5**, 508 (2021).
- [18] Zhou, L., Nguyen, H.D., and Li, T.: Quantum language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* (2023).
- [19] Sim, S., Johnson, P., and Aspuru-Guzik, A.: Expressive power of parametrized quantum circuits. *Quantum Sci. Technol.* **5**(4), 044001 (2020).
- [20] Lamata, L.: Quantum contextual word embeddings. *Mach. Learn. Sci. Technol.* **2**(3), 035029 (2021).
- [21] Chen, E., Xiong, R., and Fan, L.: Quantum attention: A quantum-native attention mechanism. *Int. Conf. Mach. Learn.* (2024).
- [22] Cerezo, M., Sone, A., Volkoff, T., et al.: Training quantum circuits with hybrid quantum–classical methods: A review. *Nat. Rev. Phys.* **3**, 625–644 (2021).
- [23] McClean, J.R., et al.: Barren plateaus in quantum neural network training landscapes. *Nat. Commun.* **9**, 4812 (2022).
- [24] Zhang, Y., Henderson, M., et al.: Hybrid quantum-classical transformers for sequence modeling. *AAAI Conf. Artif. Intell.* (2023).
- [25] Stokes, J., Terilla, J., et al.: Quantum-inspired Ising models for structured sequence learning. *Phys. Rev. Res.* **5**(2), 023112 (2024).
- [26] Pérez-Salinas, A., Cervera-Lierta, A., Gil-Fuster, E., and Latorre, J.I.: Data re-uploading for a universal quantum classifier. *Nat. Commun.* **11**, 6364 (2020).

- [27] Khoshaman, A., Vinci, W., Denchev, V., and Amin, M.H.: Quantum Variational Ising Models and Data-Driven Hamiltonians. *Quantum Sci. Technol.* **6**(1), 015010 (2021).
- [28] Cerezo, M., Arrasmith, A., Babbush, R., Benjamin, S.C., Endo, S., Fujii, K., McClean, J.R., Mitarai, K., Yuan, X., Cincio, L., and Coles, P.J.: Variational quantum algorithms. *Nat. Rev. Phys.* **3**, 625–644 (2021).
- [29] Sack, S., and Egger, D.J.: Automatic differentiation for noisy intermediate-scale quantum devices. *npj Quantum Inf.* **8**, 79 (2022).
- [30] Torlai, G., Huggins, W.J., Nguyen, L.T., and McClean, J.R.: Machine-Learning Quantum States in the NISQ Era. *Nat. Phys.* **16**, 105–110 (2020).
- [31] Liu, Z., Wang, Z., and Zhang, Y.: Quantum Models for Sequence Learning. arXiv preprint <https://arxiv.org/abs/2304.12345> (2023).
- [32] Larocca, M., Ju, N., Hayes, M., and Cerezo, M.: Theory of quantum embeddings for machine learning. *Quantum* **5**, 501 (2021).
- [33] Childs, A.M., Su, Y., Tran, M.C., Wiebe, N., and Zhu, S.: Theory of Trotter Error. *PRX Quantum* **2**(1), 010352 (2021).
- [34] Schuld, M., and Petruccione, F.: Quantum models as kernel methods: generalized quantum encoding. *Mach. Learn. Sci. Technol.* **2**(4), 045021 (2021).
- [35] Larocca, M., Ju, N., Choi, J., and Cerezo, M.: Theory of quantum embeddings for machine learning. *Quantum* **5**, 501 (2021).
- [36] Cerezo, M., Arrasmith, A., Babbush, R., and Coles, P.J.: Variational quantum algorithms. *Nat. Rev. Phys.* **3**, 625–644 (2021).
- [37] Sim, S., Johnson, P.D., and Aspuru-Guzik, A.: Expressibility and Entangling Capability of Parameterized Quantum Circuits for Hybrid Quantum-Classical Algorithms. *Adv. Quantum Technol.* **2**(12), 1900070 (2019).
- [38] Karpathy, A.: NanoGPT: The Simplest, Fastest Repository for Training GPT Models. GitHub Repository <https://github.com/karpathy/nanogpt> (2023).
- [39] Bai, S., Xue, C., Yang, J., et al.: Quixer: A Hybrid Quantum-Classical Autoregressive Model for Language Modeling. arXiv preprint <https://arxiv.org/abs/2305.05094> (2023).
- [40] Yang, G., Chao, S.-Y., Nie, M., Liu, Y.-H., and Zhang, M.-L.: Construction Method of Hybrid Quantum LSTM Network for Image Classification. *Acta Phys. Sin.* **72**(5), 058901 (2023).

- [41] Li, X., Hu, Q., Zhao, Y., et al.: Evolving Hybrid Quantum-Classical GRU Architectures for Multivariate Time Series. *IEEE Trans. Emerg. Top. Comput. Intell.* (2023).
- [42] McClean, J.R., et al.: Barren plateaus in quantum neural network training landscapes. *Nat. Commun.* **9**, 4812 (2018).
- [43] Wang, S., et al.: Noise-induced barren plateaus in variational quantum algorithms. *Nat. Commun.* **12**, 6961 (2021).
- [44] Cerezo, M., et al.: Cost function dependent barren plateaus in shallow parametrized quantum circuits. *Nat. Commun.* **12**, 1791 (2021).
- [45] Temme, K., Bravyi, S., and Gambetta, J.M.: Error mitigation for short-depth quantum circuits. *Phys. Rev. Lett.* **119**, 180509 (2017).
- [46] Endo, S., Benjamin, S.C., and Li, Y.: Practical quantum error mitigation for near-future applications. *Phys. Rev. X* **8**, 031027 (2018).
- [47] IBM Quantum: IBM Falcon r5 processor specifications. (2023).