Learning Quantum States with Deep Vision Transformers

Anuj Apte
apteanuj@uchicago.edu
University of Chicago
Chicago, IL, USA

Davi B. Costa
davicosta@uchicago.edu
University of Chicago
Chicago, IL, USA

ABSTRACT

The exponential complexity of the quantum many-body problem is a central challenge in theoretical physics and quantum information. Neural Quantum States (NQS) have emerged as a powerful variational approach, representing quantum wave functions with neural networks. Despite significant progress, capturing long-range quantum correlations and critical phenomena remains nontrivial. In this work, we propose and investigate the use of Deep Vision Transformers (ViTs) as an architecture for NQS. Focusing on the one-dimensional transverse field Ising model (TFIM), we train NQS across various couplings. By combining training accelerated with the stochastic reconfiguration technique with transfer learning across the coupling, we achieve highly accurate approximations of the ground state. Crucially, our method captures the phase transition and critical behavior expected of the TFIM, extracting critical exponents consistent with theoretical predictions. Our results indicate that ViTs, with their global attention mechanisms, are well suited to model long-range quantum correlations.

1 INTRODUCTION

Understanding quantum many-body systems remains a fundamental challenge in theoretical physics and quantum information science. These systems involve an exponentially large configuration space, and states are characterized by a complex amplitude (or wave functions) $\psi(\vec{\sigma}) = \langle\vec{\sigma}|\psi\rangle$ for each spin configuration $\vec{\sigma}$. Computing the evolution of these amplitudes involves solving the eigenvalue problem for the Hamiltonian matrix, which, for a two-dimensional qubit system, has dimensions $2^N \times 2^N$, where N represents the number of degrees of freedom. This becomes computationally intractable for systems of even moderate size, making it essential to explore approximate methods for representing quantum states.

The quantum many-body problem can be viewed as a task of efficiently compressing the exponentially large set of amplitudes into a parameterized representation. Variational approaches achieve this by introducing a functional form $\psi(\vec{\sigma};\theta) = \langle\vec{\sigma}|\psi(\theta)\rangle$ for the wave function, defined by parameters θ. The ground state can be found by optimizing θ to minimize the energy functional:

$$E(\psi(\theta)) = \frac{\langle\psi(\theta)|H|\psi(\theta)\rangle}{\|\psi(\theta)\|^2}\,. \tag{1}$$

Recent advances in machine learning have introduced Neural Quantum States (NQS) [1] as a powerful variational framework for representing quantum wave functions. NQS uses neural networks to parametrize $\psi(\vec{\sigma};\theta)$ having θ as the weights, leveraging their remarkable expressivity to encode high-dimensional quantum correlations. Initial successes employed Restricted Boltzmann Machines (RBMs) [1], which inspired research into more complex architectures, such as Convolutional Neural Networks (CNNs) [2], and Transformers [3].

Despite these advances, capturing long-range correlations and critical phenomena remains challenging for traditional architectures. Transformers, with their global attention mechanisms, offer a promising alternative. In particular, Viteritti et al. [3] introduced the Vision Transformer (ViT) as a neural variational wave function, demonstrating its ability to model frustrated quantum spin systems. The ViT architecture uses self-attention to encode local and global correlations effectively, followed by a single complex feedforward neural network (FFNN) layer to output the complex amplitudes of the wave function.

1.1 Contributions

Building on the work of Viteritti et al. [3], we analyze the application of ViTs to Neural Quantum States (NQS) for studying quantum many-body systems. Our contributions include:

(1) Demonstrating the effectiveness of ViTs in capturing long-range correlations and critical behavior in the transverse field Ising model (TFIM) [4].
(2) Utilizing transfer learning techniques to significantly reduce computational overhead when exploring the phase diagram of the TFIM.
(3) Extracting universal scaling exponents near criticality, validating the ViT-based NQS approach against theoretical predictions.

The rest of this paper is organized as follows: Section 2 describes the many-body problem with TFIM as a benchmark system, and highlights its critical phenomena. Section 3 introduces the theoretical framework for NQS and details the ViT architecture. Section 4 presents our numerical experiments, focusing on energy convergence and critical scaling. Finally, Section 5 discusses the implications of this work and future directions.

2 THE TRANSVERSE FIELD ISING MODEL

2.1 Model

Quantum spin systems are widely used mathematical frameworks for modeling physical phenomena in condensed matter physics and quantum information theory [5]. These systems are applied to study topics such as magnetism and quantum circuits. They consist of a discrete set of sites, often corresponding to the vertices of a lattice, with each site associated with a finite-dimensional Hilbert space. For instance, the Hilbert space at a given site might represent a particle with spin-s fixed at that lattice vertex. We will study a linear chain of spins each having a two-dimensional Hilbert space with the basis elements corresponding to the spin pointing up or down. The transverse field Ising model (TFIM), is defined by the Hamiltonian:

$$\hat{H} = -J\sum_i Z_i Z_{i+1} - g\sum_i X_i, \tag{2}$$

where $J > 0$ is the ferromagnetic coupling between neighboring spins, g is the strength of the transverse field, Z_i and X_i are Pauli matrices representing the spin operators in the z- and x-directions. By changing the units of energy, we can set $J = 1$. We choose periodic boundary conditions such that spins are on a ring with the last spin being connected to the first spin.

This model captures the interplay between competing interactions: the first term favors alignment of spins along the z-direction (ferromagnetic order), while the second term induces quantum fluctuations by driving spins to align along the x-direction. The competition between these terms leads to rich physics, including a quantum phase transition.

For a system of N spins, the Hilbert space grows exponentially as 2^N, rendering exact solutions infeasible for large N. Instead, approximate numerical techniques, including variational methods and neural network quantum states (NQS), have proven powerful for studying such systems.

2.2 Phases of the TFIM

The TFIM exhibits two distinct phases at zero temperature:

(1) **Ferromagnetic Phase:** For small transverse field strengths ($g \ll 1$), the system minimizes its energy by aligning all spins in the z-direction, resulting in a magnetically ordered state. The ground state has long-range correlations and a nonzero magnetization in the z-direction.
(2) **Paramagnetic Phase:** For large transverse field strengths ($g \gg 1$), the system minimizes its energy by aligning spins in the x-direction, breaking the long-range order. The ground state is dominated by quantum fluctuations and exhibits no net magnetization.

These phases are separated by a critical point at g_c, where the system undergoes a second-order quantum phase transition. The critical behavior is governed by universal scaling laws, which are a hallmark of second-order phase transitions [6].

2.3 Phase Transition and Scale Invariance

The quantum phase transition in the TFIM occurs at a critical value of the transverse field, $g_c \simeq 1$ (in 1D). The correlation length diverges as the system approaches the critical point, following the relation $\xi = \xi_0 |g - g_c|^{-\nu}$, where $\nu > 0$ is a positive critical exponent. For systems of finite size, L, finite-size scaling theory explains how observables depend on the system size L and the distance from the critical point [7]. A central idea of this theory is that near the critical coupling g_c, the correlation length ξ becomes comparable to the system size L. Consequently, the microscopic length scale, determined by the lattice spacing and governing the range of interactions, no longer influences correlation functions on length scales larger than the lattice spacing. As a result, the system exhibits scale invariance: properties of the system remain unchanged under the rescaling of length and energy scales.

This behavior is characterized by critical exponents, which determine how various physical observables such as magnetization behave near the transition:

$$\langle M \rangle \propto |g - g_c|^{\beta} . \tag{3}$$

In contrast to the magnetization, we choose a more numerically robust observable which is the two-point correlation function $G(r) = \langle Z_i Z_{i+r} \rangle$. We choose $r = L/2$, which corresponds to diametrically opposite points on the ring ensuring that the points are as far apart as they can be. This correlation function in the vicinity of the critical point behaves as:

$$\langle Z_i Z_{i+L/2} \rangle = L^{-2\beta/\nu} F\big(L^{1/\nu}(g - g_c)\big) . \tag{4}$$

Thus, if we plot $L^{2\beta/\nu} \langle Z_i Z_{i+L/2} \rangle$ against $L^{1/\nu}(g - g_c)$ for appropriate choice of g_c, β, ν the curves for different values of L should collapse onto a single curve. In 1D, the TFIM belongs to the same universality class as the 2D classical Ising model, with critical exponents $\beta = 1/8$ and $\nu = 1$. These exponents reflect the universal nature of the transition and are independent of microscopic details [8].

3 VISION TRANSFORMER NEURAL QUANTUM STATES

3.1 Neural Quantum States

Neural Quantum States (NQS) represent a powerful variational ansatz for many-body quantum systems, where the wavefunction is parameterized by an artificial neural network. The fundamental idea is to express the quantum state as a complex-valued function:

$$\psi(\sigma; \theta) = \langle \sigma | \psi(\theta) \rangle , \tag{5}$$

where σ represents a basis state (typically in the computational basis), and θ denotes the network parameters. This neural network takes as input a configuration σ (e.g., spin configurations, occupation numbers) and outputs both amplitude and phase of the wavefunction.

The power of this approach lies in its flexibility and expressivity. The same architectural framework can be adapted to diverse quantum systems by choosing appropriate symmetries and network architectures. Applications span across multiple domains of quantum physics, from strongly correlated electrons in condensed matter [9] molecular systems in quantum chemistry [10], and even to lattice gauge theories in high-energy physics [11].

The variational Monte Carlo (VMC) framework with NQS leverages the universal approximation capabilities of neural networks while maintaining quantum mechanical properties. For instance, translational invariance can be encoded through convolutional architectures, while phase relationships can be preserved through complex-valued networks or separate networks for amplitude and phase.

3.2 Expectation Values and Gradients

The computation of quantum expectation values forms the core of the NQS framework. For an operator $\hat{O}$, the expectation value is given by:

$$\langle \hat{O} \rangle = \frac{\langle \psi(\theta) | \hat{O} | \psi(\theta) \rangle}{\langle \psi(\theta) | \psi(\theta) \rangle} = \sum_{\sigma} P(\sigma) O_{\text{loc}}(\sigma) , \tag{6}$$

where $P(\sigma) = |\psi(\sigma; \theta)|^2 / \|\psi(\theta)\|^2$ is the probability distribution and

$$O_{\text{loc}}(\sigma) = \sum_{\sigma'} \frac{\langle \sigma | \hat{O} | \sigma' \rangle \psi(\sigma'; \theta)}{\psi(\sigma; \theta)} , \tag{7}$$

is the local estimator.

These quantities are estimated through Markov Chain Monte Carlo (MCMC) sampling from $P(\sigma)$. The energy gradient components required for optimization are:

$$\nabla E = \frac{\partial E}{\partial \theta_i} = 2\mathrm{Re}\left[\langle E_{\mathrm{loc}} O_i^* \rangle - \langle E_{\mathrm{loc}} \rangle \langle O_i^* \rangle\right] , \tag{8}$$

where $O_i = \partial_{\theta_i} \log \psi(\sigma;\theta)$ are the logarithmic derivatives. We use the NetKet package for computing expectation values, gradients and training the networks [12].

Notably, the Hamiltonian expectation value can be estimated accurately with fewer samples compared to other observables due to the zero-variance principle: as the wavefunction approaches an eigenstate, the variance of the local energy vanishes, leading to more efficient sampling.

3.3 Stochastic Reconfiguration Method

The Stochastic Reconfiguration (SR) method, equivalent to the quantum natural gradient, provides an efficient optimization strategy for NQS [13]. The parameter updates follow:

$$\theta^{(t+1)} = \theta^{(t)} - \eta S^{-1} \nabla E \, , \tag{9}$$

where η is the learning rate at step t and S is the quantum Fisher information matrix (SR matrix):

$$S_{ij} = \langle O_i^* O_j \rangle - \langle O_i^* \rangle \langle O_j \rangle \, . \tag{10}$$

The SR matrix captures the local geometry of the quantum state manifold, enabling more efficient optimization compared to standard gradient descent. However, the SR matrix often contains near-zero eigenvalues corresponding to flat directions in parameter space, which can lead to numerical instabilities.

To address this, a diagonal shift is introduced:

$$S \longrightarrow S + \lambda I \tag{11}$$

where λ is a regularization parameter. This modification effectively interpolates between natural gradient ($\lambda \to 0$) and standard gradient descent ($\lambda \to \infty$). Stochastic reconfiguration was implemented following the algorithm outlined in [14, 15] with a diagonal shift of 10^{-4}. This algorithm allows one to compute the weight update without inverting a $N_p \times N_p$ sized matrix where N_p is the number of parameter. By using a linear algebra identity, one can instead invert a matrix of size $N_s \times N_s$ where N_s is the number of samples. Therefore the inverting the SR matrix takes a fixed amount of time and memory independent of the size of the neural network.

We used a cosine decay learning schedule with initial learning rate set to $\eta_0 = 0.1$. In general the learning rates used for NQS with SGD tend to be higher than the ones used for conventional deep learning applications. The SR method accelerates convergence by accounting for the non-trivial geometry of quantum state manifolds. When parameters affect the wavefunction similarly (leading to flat directions), the SR matrix identifying these correlations allows for more efficient parameter updates compared to naive gradient descent.

3.4 Vision Transformer Architecture

Vision Transformers (ViTs) [16] adapt the transformer architecture [17], originally developed for natural language processing, to process image data. Unlike traditional convolutional neural networks (CNNs) that rely on localized receptive fields, ViTs treat an image as a sequence of fixed-size patches, similar to words in a sentence. Each patch is flattened into a vector and projected into an embedding space, forming the input sequence for the transformer.

The architecture consists of several key components: a positional encoding layer to inject spatial information into the patch embeddings, multiple transformer layers for capturing global dependencies through self-attention, and a classification head for downstream tasks. The self-attention mechanism enables ViTs to model long-range relationships across patches, making them particularly effective for understanding high-level features. The architecture that we use here was introduced in [3], which adapts the ViT to the task of predicting the wave-function of the spin system.

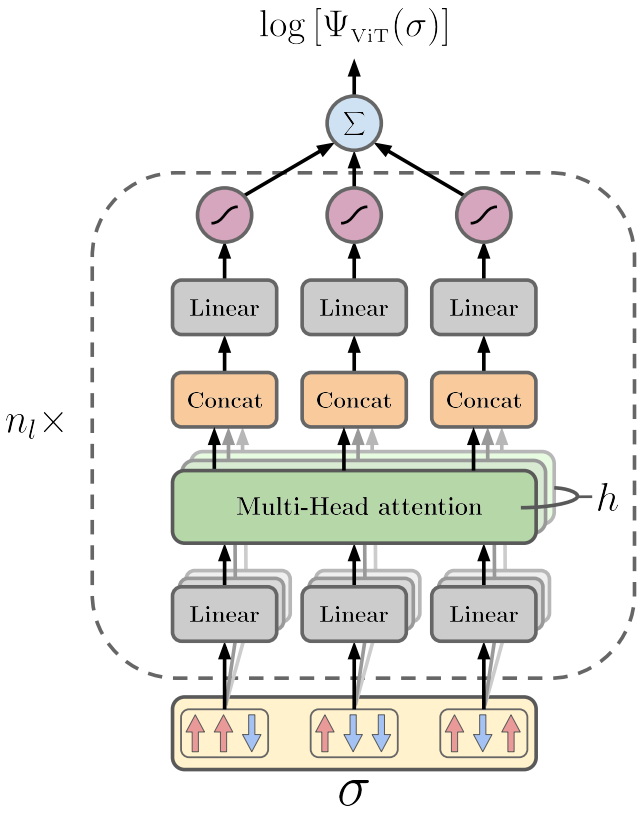

Figure 1: The architecture of Vision Transformer adapted to neural quantum states. The ViT output serves as an effective feature extractor before being fed into a feed forward layer for the output. Figured adpated from [3].

The output of the neural quantum is $\log \psi_\theta(\sigma)$ where θ are the weights of the network, and σ is a spin configurations in the Hilbert space $\sigma = (\sigma_1, \ldots, \sigma_L)$. Starting from a spin configuration $\sigma = (\sigma_1, \ldots, \sigma_L)$, we split it into N patches of b elements: $\boldsymbol{x}_i = (\sigma_{(i-1)b+1}, \ldots, \sigma_{(i-1)b+b})$, for $i = 1, \ldots, N$ (the total number

of sites must be a multiple of b). The sequence of these patches is then used to compute the attention vectors. Based on the physical principle that the interactions between spins depend on relative positions and not on details of the configuration, we can simplify the attention so that they depend on positions i and j, but not on the actual values of the spins in these patches leading to:

$$\boldsymbol{A}_i^{\mu} = \sum_{j=1}^{N} \alpha_{ij}^{\mu} V^{\mu} \boldsymbol{x}_j, \tag{12}$$

where V^{μ} is a $r \times b$ matrix with $r = d/h$, and d is the so-called *embedding dimension* that must be a multiple of the number of heads h [18].

Since the network output is a single complex number, we modify the standard ViT architecture by treating it's output before the classification head as a feature extractor. This output is then fed to a single feed-forward layer with complex weights so that the output is of the overall network is complex. The d-dimensional hidden representation is obtained as $\boldsymbol{z} = \sum_{i=1}^{n} \boldsymbol{y}_i \in \mathbb{R}^d$ at the end of the ViT block. The final mapping is performed by an output layer parametrized as a shallow network, namely $\mathrm{Log}[\Psi_\theta(\sigma)] = \sum_{\alpha=1}^{d} g(b_\alpha + \boldsymbol{w}_\alpha \cdot \boldsymbol{z})$, with non-linearity $g(\cdot) = \mathrm{logcosh}(\cdot)$ and complex-valued trainable parameters $\{b_\alpha, \boldsymbol{w}_\alpha\}_{\alpha=1}^{d}$ [3]. The network architecture is visualized in Figure 1.

Note that for our investigation we had spin chains of length $L = 12, 16, 20$, so we chose a patch size of $b = 4$. The number of heads was equal to $h = 4$, while we picked a single layer $n_l = 1$ with an embedding dimension of $d = 32$. Adding additional layers did not improve the performance of the model.

4 RESULTS

4.1 Energy

To achieve highly accurate results, we leveraged transfer learning as a key technique in our training process. Specifically, we initialized the model using weights obtained from training at a coupling value close to the target coupling [19]. This approach significantly accelerated convergence since the model already possessed a reasonably good approximation of the wave function from the nearby coupling. As a result, we were able to obtain energies with a relative error of 10^{-6} after just a few hundred optimization steps as shown in Figure 2.

The use of transfer learning is particularly advantageous in systems with continuous parameter spaces, such as coupling strengths, as it allows the model to reuse and refine learned features from similar configurations. This reduces the number of Monte Carlo samples and optimization iterations required to reach high precision, making the process computationally efficient.

Let us now consider the two regimes of $g \gg 1$ and $g \ll 1$ and understand how energy should behave as a function of g. When $g \gg 1$, in the paramagnetic disordered phase the spins are each in an X eigenstate with eigenvalue -1 and thus the energy per spin is simply $-g$. However, when $g \ll 1$ in the ferromagnetic ordered phase all the spins are each in Z eigenstate with eigenvalues either $+1$ or -1. Note that the fact that all spins can take one of the two values, corresponds to symmetry breaking of the $\mathbb{Z}_2$ symmetry in the Ising model [20].

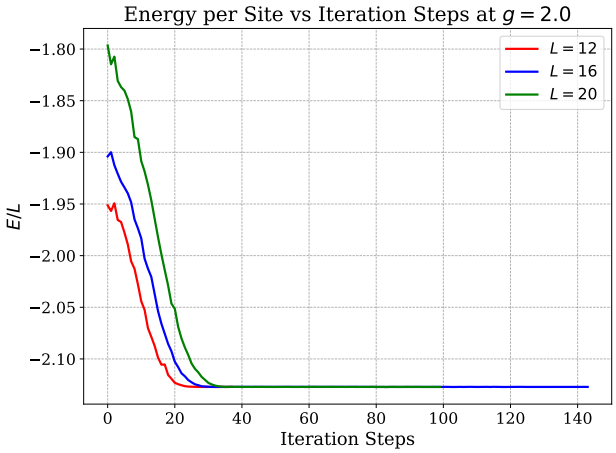

Figure 2: Convergence to ground state at $g = 2.0$ for Vision Transformer neural quantum state.

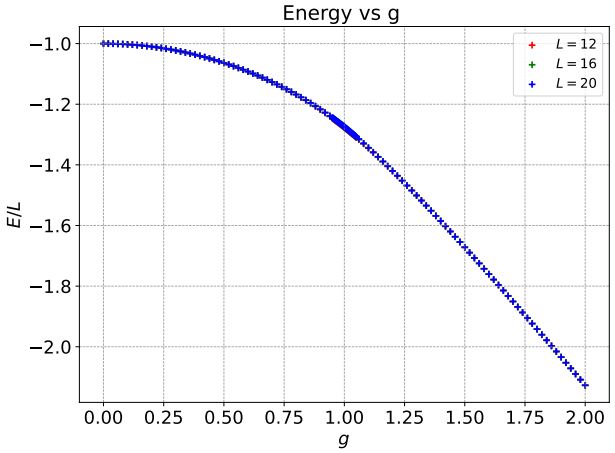

Figure 3: Energy as a function of coupling g for 1D Tranverse field Ising model.

Thus, the energy per spin in this regime will be -1. The linear behavior for $g \gg 1$, and energy saturating to -1 per spin can be observed in Figure 3. Note that since the system sizes are large enough, the curves for the different system sizes lie exactly on top of each other indicating that the systems are a good approximation to the continuum behavior.

Since the ground state is the lowest energy eigenstate of the Hamiltonian, the variance for the ground state vanishes. Thus, we can compute the variance of the Hamiltonian during the Markov chain Monte-Carlo sampling during the training and use it as a way to understand how close we are to the ground state without previous knowledge of the ground state energy. The variance observed is 10^{-4} or smaller for all the couplings, indicating excellent convergence to the ground state as shown in Figure 4.

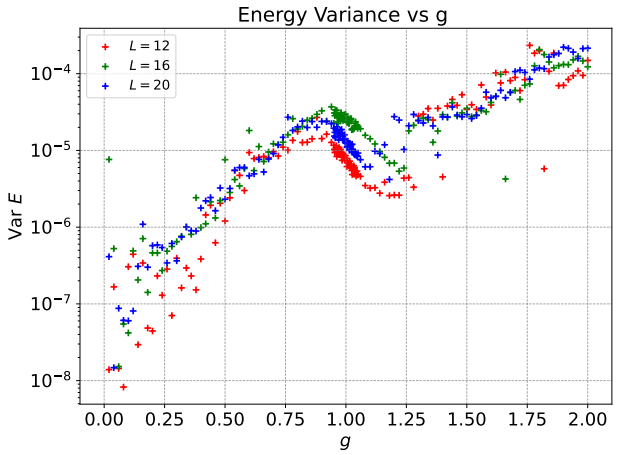

Figure 4: Variance of energy as a function of coupling g for 1D Tranverse field Ising model.

4.2 Correlation Function and Phase Transition

We now turn our attention to the computation of the two-point correlation function for diametrically opposite points $\langle Z_i Z_{i+L/2} \rangle$. After the ground states have been trained, computing the expectation correlation function gives insight into whether the model is in the paramagnetic or the ferromagnetic state and allows us to study the second-order continuous phase transition. When $g \gg 1$ the spins are each in an X eigenstate and thus the correlation function vanishes. Whereas, when $g \ll 1$ the spins are each in an Z eigenstate and thus the correlation function equals unity. Figure 5 shows how the correlation function changes as a function of the coupling from 1 to 0 as the coupling g is increased. Note that the correlation function drops sharply near the vicinity of the phase transition at $g_c = 1$ as the model goes from the ferromagnetic to the paramagnetic phase.

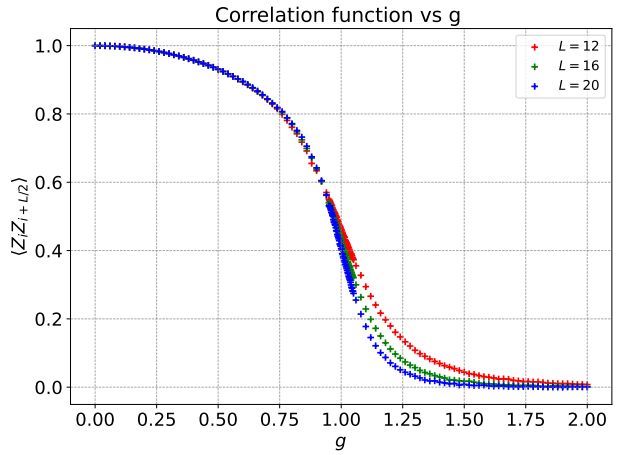

Figure 5: $\langle Z_i Z_{i+L/2} \rangle$ two-point correlation function as a function of coupling g for 1D Tranverse field Ising model.

The second-order phase transition at $g_c = 1$ is analyzed by training states in increments of 0.002 from $0.95 - 1.05$ so that we can carefully study how the correlation function behaves as a function of coupling. Based on the discussion in Section 2, by plotting $L^{2\beta/\nu} \langle Z_i Z_{i+L/2} \rangle$ against $L^{1/\nu}(g - g_c)$ and adjusting g_c, β, ν such that the curves for different values of L collapse onto a single curve, we can numerically obtain the values for the critical exponents. For more details of the numerical procedure consult Appendix B of [19] or [21]. By statistical bootstrapping [22], we compute the uncertainties in our estimate for g_c, β, ν. The values obtained by following this procedure are:

$$g_c = 0.998 \pm 0.002, \quad \beta = 0.124 \pm 0.001, \quad \nu = 1.000 \pm 0.001. \quad (13)$$

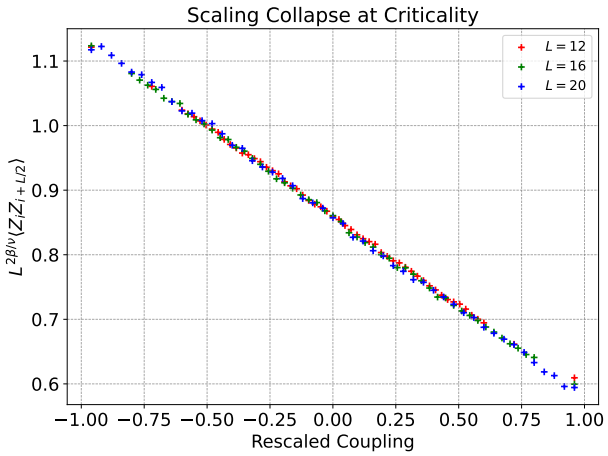

Figure 6: $\langle Z_i Z_{i+L/2} \rangle$ two-point correlation function curve collapse in the vicinity of the critical point $g_c = 0.998 \pm 0.002$.

These values precisely match the well-known 2D classical Ising model critical exponents. The scaling collapse is plotted in Figure 6, providing a visual confirmation of the second-order nature of the phase transition.

5 CONCLUSION

This work demonstrates the efficacy of Vision Transformers (ViTs) in quantum many-body physics, specifically in encoding quantum correlations and computing ground states of complex systems. Our findings reveal several significant advantages of the ViT architecture in this domain. The global attention mechanism proves particularly powerful in simultaneously capturing both local interactions and long-range quantum correlations, offering a natural framework for modeling quantum systems. Furthermore, our implementation of transfer learning between different couplings substantially reduces the computational overhead, enabling efficient exploration of the phase diagram.

The success of ViTs in this context stems from their inherent ability to process global information through self-attention mechanisms. Unlike traditional convolutional approaches, which primarily focus on local features, ViTs can naturally adapt to the non-local nature of quantum correlations. This capability becomes especially

crucial when dealing with strongly correlated quantum systems where long-range interactions play a fundamental role.

However, two main limitations warrant discussion. The computational complexity of our approach increases significantly with lattice size, presenting challenges for scaling to larger systems. While we implement Stochastic reconfiguration to accelerate training convergence, this optimization introduces additional stability concerns that require careful parameter tuning.

Looking ahead, several promising directions emerge for future research:

- **Systematic Ablation Studies:** A comprehensive investigation of individual ViT components' contributions to model performance and training convergence is essential. This analysis would provide valuable insights for optimizing the architecture for specific quantum systems.
- **Attention Map Analysis:** Detailed examination of the learned attention patterns could reveal fundamental insights into how the model encodes quantum correlations. This understanding could lead to more efficient architectural designs.
- **Higher-Dimensional Extensions:** Extending our approach to 2D and 3D lattices represents a natural progression [9]. This extension will require careful consideration of lattice symmetry constraints and computational efficiency, potentially necessitating new architectural modifications.

Our results suggest that transformer-based architectures hold significant promise for quantum many-body physics, potentially offering a new paradigm for studying complex quantum systems. The ability to efficiently compute ground states while naturally incorporating both short and long-range correlations positions this approach as a valuable tool in the computational physics toolkit. Future developments addressing the identified limitations could further expand the applicability and impact of this method across various quantum systems.

REFERENCES

[1] Giuseppe Carleo and Matthias Troyer. Solving the quantum many-body problem with artificial neural networks. *Science*, 355(6325):602–606, 2017.

[2] Kenny Choo, Titus Neupert, and Giuseppe Carleo. Two-dimensional frustrated J_1-J_2 model studied with neural network quantum states. *Physical Review B*, 100:125124, September 2019.

[3] Luciano Loris Viteritti, Riccardo Rende, and Federico Becca. Transformer variational wave functions for frustrated quantum spin systems. *Physical Review Letters*, 130(23):236401, June 2023.

[4] Eduardo Fradkin. *Field Theories of Condensed Matter Physics*. Cambridge University Press, February 2013.

[5] Amanda Young. Quantum spin systems, 2023.

[6] H. Eugene Stanley. Scaling, universality, and renormalization: Three pillars of modern critical phenomena. *Rev. Mod. Phys.*, 71(2):S358–S366, 1999.

[7] Michael E. Fisher and Michael N. Barber. Scaling theory for finite-size effects in the critical region. *Phys. Rev. Lett.*, 28:1516–1519, 1972.

[8] Gesualdo Delfino and Jacopo Viti. Potts q-color field theory and scaling random cluster model. *Nuclear Physics B*, 852(1):149–173, November 2011.

[9] K. Choo, T. Neupert, and G. Carleo. Two-dimensional frustrated J_1-J_2 model studied with neural network quantum states. *Phys. Rev. B*, 100:125124, Sep 2019.

[10] Ingrid von Glehn, James S. Spencer, and David Pfau. A self-attention ansatz for ab-initio quantum chemistry, 2023.

[11] D. Luo, Z. Chen, J. Carrasquilla, and B.K. Clark. Autoregressive neural network for simulating open quantum systems via a probabilistic formulation. *Phys. Rev. Lett.*, 128:090501, Feb 2022.

[12] Filippo Vicentini et al. NetKet 3: Machine Learning Toolbox for Many-Body Quantum Systems. *SciPost Phys. Codeb.*, 2022:7, 2022.

[13] Sandro Sorella. Green function Monte Carlo with stochastic reconfiguration. *Phys. Rev. Lett.*, 80(20):4558–4561, 1998.

[14] Ao Chen and Markus Heyl. Efficient optimization of deep neural quantum states toward machine precision. 2023.

[15] Riccardo Rende, Luciano Loris Viteritti, Lorenzo Bardone, Federico Becca, and Sebastian Goldt. A simple linear algebra identity to optimize large-scale neural network quantum states. 2023.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, Jun 2021.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, Dec 2017.

[18] Riccardo Rende and Luciano Loris Viteritti. Are queries and keys always relevant? a case study on transformer wave functions, 2024.

[19] Anuj Apte, Clay Córdova, Tzu-Chen Huang, and Anthony Ashmore. Deep learning lattice gauge theories. *Phys. Rev. B*, 110:165133, Oct 2024.

[20] J. A. Ringler, A. I. Kolesnikov, and K. A. Ross. Single-ion properties of the transverse-field ising model material $conb_2o_6$. *Phys. Rev. B*, 105:224421, Jun 2022.

[21] Somendra M Bhattacharjee and Flavio Seno. A measure of data collapse for scaling. *J. Phys. A: Math. Gen.*, 34(33):6375–6380, 2001.

[22] Bradley Efron. Bootstrap methods: Another look at the jackknife. In *Breakthroughs in Statistics: Methodology and Distribution*, pages 569–593. Springer New York, 1992.