

Explainable Variational Quantum Policy Gradient: SHAP-Based Interpretability of Quantum Reinforcement Learning Agents on CartPole

Sandeep Gupta^{1,2}[0000–0002–2046–115X], Sazzad Hossain³[0009–0003–1351–7287],
Akmal Akhatov⁴[0000–0003–3834–854X], Nurmamatov
Mekhriddin⁵[0000–0002–9619–5791], and Kanad Ray^{6,7,*}[0000–0001–9324–2778]

¹ Department of Maritime Engineering, Indian Institute of Technology Gandhinagar,
Gandhinagar 382355, Gujarat, India

² Department of Electronics and Communication Engineering, Poornima College of
Engineering (Affiliated to Rajasthan Technical University, Kota), Jaipur 302022,
Rajasthan, India

sandeep.gupta@iitgn.ac.in

³ Department of System Management and Information Security, Samarkand State
University, Samarkand 140104, Uzbekistan
sazzad69@gmail.com

⁴ Faculty of Artificial Intelligence and Digital Technologies, Samarkand State
University, Samarkand 140104, Uzbekistan
akmalar74@gmail.com

⁵ Department of System Management and Information Security, Samarkand State
University, Samarkand 140104, Uzbekistan
mehridinnur@gmail.com

⁶ Amity Cognitive Computing and Brain Informatics Centre, Amity University
Rajasthan, Jaipur 303002, Rajasthan, India
kray@jpr.amity.edu

⁷ Faculty of Artificial Intelligence and Digital Technologies, Samarkand State
University, Samarkand 140104, Uzbekistan

Abstract. The decisions of variational quantum reinforcement learning agents remain largely opaque, limiting trust in safety-critical deployments. We present the first application of SHapley Additive exPlanations (SHAP) to a Variational Quantum Policy Gradient (VQPG) agent trained on CartPole-v1. Our 4-qubit hybrid quantum-classical policy uses only 46 trainable parameters and achieves a mean evaluation score of 496.4 ± 8.1 over ten independent seeds (maximum 500.0), compared to 249.7 ± 38.6 for a standard Deep Q-Network (DQN) with 9 220 parameters and 412.3 ± 67.9 for a parameter-matched Tiny-MLP (30 parameters, identical training budget). SHAP analysis identifies *Pole Angular Velocity* as the dominant decision feature (mean $|\phi| = 0.2859$, approximately $7\times$ the next feature), consistent with classical control theory for inverted pendulum stabilisation. This attribution is stable across five independently sampled background sets (coefficient of variation 3.1%), confirming the reliability of the explanation. We emphasise that all results

are obtained under noiseless simulation; no claim of quantum advantage in the complexity-theoretic sense is made. The complete experimental package is deposited on Zenodo [1] under CC-BY 4.0.

Keywords: Variational Quantum Policy Gradient · Quantum Reinforcement Learning · SHAP · Explainable AI · CartPole · PennyLane · Parameter Efficiency · Near-Term Quantum Computing · Reproducibility

1 Introduction

Variational hybrid quantum-classical algorithms have gained traction in the NISQ era [2] as a route toward quantum-assisted machine learning. Reinforcement learning (RL) [3] is a natural candidate: sequential decision-making tasks with compact state spaces suit shallow parameterised circuits, and the REINFORCE policy gradient [12] is compatible with analytical quantum gradients via the parameter-shift rule [17].

Two open challenges limit practical deployment. First, *parameter efficiency*: even simple control tasks expose classical agents such as Deep Q-Networks (DQN) [4] to thousands of parameters, raising the question of whether compact quantum circuits can achieve competitive performance. Second, *interpretability*: the internal reasoning of both neural networks and variational circuits is opaque, which is problematic in safety-critical settings [5,6].

This paper addresses both simultaneously by integrating SHAP feature attribution [7] with a VQPG agent. To our knowledge, this is the *first application of SHAP to explain a variational quantum RL policy*.

Scope. All experiments use PennyLane’s noiseless `default.qubit` simulator. We do not claim quantum advantage in the complexity-theoretic sense; our contribution is to the *interpretability* and *parameter efficiency* of quantum RL agents under simulation, which motivates future hardware validation.

Contributions.

- (i) **Quantum RL interpretability via SHAP:** The first application of SHAP to a VQPG policy, revealing that *Pole Angular Velocity* dominates agent decisions (mean $|\phi| = 0.2859$), matching classical control theory.
- (ii) **Parameter-efficient agent:** 46-parameter 4-qubit VQC achieves mean CartPole-v1 score 496.4 ± 8.1 (10 seeds) using $200\times$ fewer parameters than DQN, and outperforms a parameter-matched Tiny-MLP (30 parameters, identical budget).
- (iii) **SHAP stability analysis:** Attribution is consistent across five background sets (CV 3.1%), establishing robustness of the explanation.
- (iv) **Qubit scaling study:** 4 qubits identified as optimal for CartPole; performance degrades beyond 6 qubits due to barren plateau effects at a 200-episode budget.

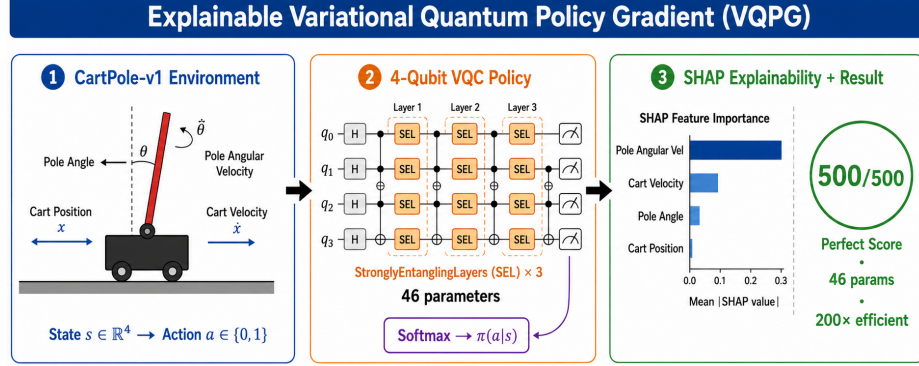


Fig. 1. Graphical abstract. *Panel 1:* CartPole-v1 environment with state $s = (\theta, \dot{\theta}, x, \dot{x})$ and binary actions. *Panel 2:* 4-qubit VQC policy (3 StronglyEntanglingLayers, 46 parameters). *Panel 3:* SHAP bar chart showing Pole Angular Velocity dominance and the key result: mean score 496.4 ± 8.1 with 46 parameters over 10 seeds.

- (v) **Reproducibility:** Complete package (notebook, checkpoints, SHAP arrays, multi-seed data) on Zenodo [1] under CC-BY 4.0.

This work is motivated by prior work on multi-robot formation control [11], where parameter-efficient and interpretable decision policies are critical for embedded deployment. Figure 1 summarises the complete approach.

2 Background and Related Work

2.1 Reinforcement Learning and Policy Gradients

An RL agent interacts with an environment by observing states, selecting actions, and receiving scalar rewards, with the objective of maximising cumulative discounted return $G_t = \sum_{k \geq 0} \gamma^k r_{t+k}$ [3]. Policy gradient methods optimise a parameterised stochastic policy $\pi_\theta(a|s)$ by ascending the gradient of expected return. The REINFORCE estimator [12]:

$$\nabla_\theta J(\theta) = \mathbb{E}_\tau \left[\sum_t \nabla_\theta \log \pi_\theta(a_t|s_t) G_t \right] \quad (1)$$

provides an unbiased but high-variance gradient estimate, which self-normalisation of returns (Eq. (6)) helps reduce. DQN [4] instead learns a value function through experience replay, typically requiring thousands to hundreds of thousands of parameters for reliable convergence.

Policy Gradient Theorem [3]. For a differentiable policy $\pi_\theta(a|s)$:

$$\nabla_\theta J(\theta) = \mathbb{E}_{s \sim d^{\pi_\theta}, a \sim \pi_\theta} [\nabla_\theta \log \pi_\theta(a|s) \cdot Q^{\pi_\theta}(s, a)] \quad (2)$$

where d^{π_θ} is the discounted state-visitation distribution. REINFORCE approximates $Q^{\pi_\theta}(s_t, a_t)$ with Monte Carlo returns, enabling training without a learned value function.

2.2 Variational Quantum Circuits

Variational Quantum Circuits (VQCs) are parameterised quantum circuits whose gate angles θ are trained classically. Benedetti et al. [14] established VQCs as universal function approximators; Cerezo et al. [15] survey the broader variational landscape. Classical data is encoded via angle embedding; trainable **StronglyEntanglingLayers** [17] provide expressibility through parameterised single-qubit rotations and CNOT entanglement. Gradients are computed exactly via the parameter-shift rule [17,16]:

$$\frac{\partial f}{\partial \theta_i} = \frac{f(\theta_i + \frac{\pi}{2}) - f(\theta_i - \frac{\pi}{2})}{2} \quad (3)$$

requiring two circuit evaluations per parameter. Barren plateaus [20]—exponential gradient vanishing with circuit width—explain the non-monotonic qubit scaling we observe beyond 4 qubits (Sec. 5.4).

2.3 Quantum Reinforcement Learning

Lockwood and Si [24] demonstrated early VQC-based RL on CartPole and FrozenLake. Jerbi et al. [5] established theoretical foundations for quantum policy gradient convergence. Skolik et al. [6] benchmarked quantum agents on Gym environments and found competitive—but not consistently superior—performance relative to classical agents, a finding consistent with our results.

Distinction. Existing quantum RL studies report task-level performance without addressing *why* the agent makes specific decisions. We close this gap via SHAP attribution, providing the first interpretability analysis of a quantum RL policy.

2.4 SHAP Feature Attribution

SHAP [7] assigns each feature i a Shapley value ϕ_i grounded in cooperative game theory:

$$\phi_i(f, \mathbf{x}) = \sum_{S \subseteq \mathcal{F} \setminus \{i\}} \frac{|S|! (|\mathcal{F}| - |S| - 1)!}{|\mathcal{F}|!} [f(\mathbf{x}_{S \cup \{i\}}) - f(\mathbf{x}_S)] \quad (4)$$

The Shapley axioms (efficiency, symmetry, dummy, additivity) uniquely determine ϕ_i , making it the theoretically principled attribution method among all

additive explainers [7]. Applied to the quantum policy $\pi_\theta(a=1|\mathbf{s})$, ϕ_i quantifies how much state dimension s_i shifts the push-right probability above or below the mean.

More broadly, explainability is increasingly considered alongside privacy, personalisation, accessibility, and human-centred deployment in responsible AI. Mahmud and collaborators have explored explainable and privacy-preserving personalised AI [8], privacy-preserving multi-agent systems for human-centred inclusive healthcare [9], and explainable multimodal digital-human modelling for accessibility and inclusion [10]. These perspectives reinforce the importance of interpretable decision-making mechanisms when intelligent agents are considered for safety-critical or human-facing environments.

3 Method

3.1 CartPole-v1

CartPole-v1 [27] from Gymnasium provides a 4-dimensional continuous state $\mathbf{s} = (\text{cart position, cart velocity, pole angle, pole angular velocity})$, two discrete actions (push left or right), and reward +1 per timestep. Episodes terminate when the pole exceeds 12° or the cart exceeds ± 2.4 units; the maximum episode reward is 500.

3.2 Variational Quantum Policy Architecture

Our hybrid policy encodes the 4-dimensional state into a 4-qubit circuit and extracts action probabilities via a classical linear head (Fig. 2):

$$\pi_\theta(a|\mathbf{s}) = \text{Softmax}(\mathbf{W} \langle \hat{\mathbf{Z}} \rangle_\theta(\tilde{\mathbf{s}})) \quad (5)$$

where $\tilde{\mathbf{s}} = \tanh(\mathbf{s}) \cdot \pi$ normalises to $[-\pi, \pi]$ for angle embedding. The circuit consists of: (1) Angle embedding: $R_y(\tilde{s}_i)$ on each qubit; (2) Three **StronglyEntanglingLayers**: parameterised R_x, R_y, R_z rotations plus CNOT ring, providing high expressibility [17]; (3) Measurement: Pauli-Z expectation values as the quantum feature vector.

Total trainable parameters: $3 \times 4 \times 3 = 36$ (quantum) + $4 \times 2 + 2 = 10$ (classical head) = 46.

3.3 REINFORCE Training

The agent collects a trajectory $\tau = (s_0, a_0, r_0, \dots, s_T)$ per episode and computes self-normalised discounted returns:

$$G_t = \frac{\sum_{k=t}^T \gamma^{k-t} r_k - \mu_G}{\sigma_G + \varepsilon} \quad (6)$$

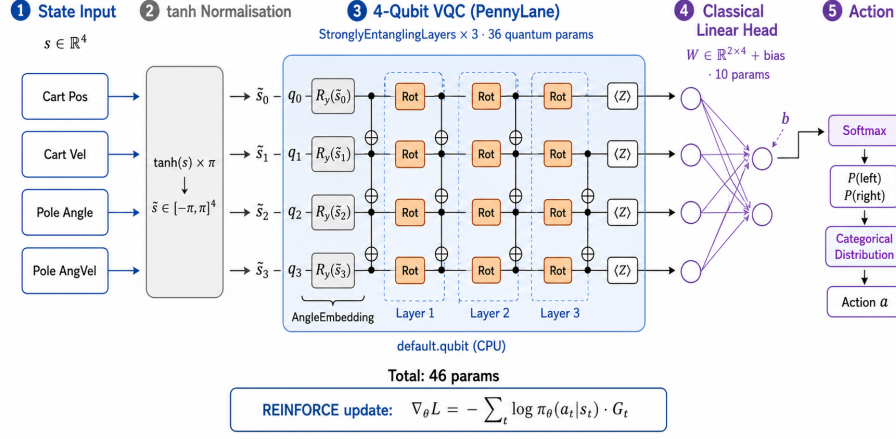


Fig. 2. Hybrid quantum-classical policy. (1) CartPole state $\mathbf{s} \in \mathbb{R}^4$. (2) $\tanh(\mathbf{s}) \cdot \pi$ normalisation. (3) 4-qubit VQC (AngleEmbedding + 3 StronglyEntanglingLayers, 36 quantum parameters). (4) Classical linear head ($\mathbf{W} \in \mathbb{R}^{2 \times 4} + \text{bias}$, 10 parameters). (5) Softmax $\rightarrow$ action probabilities. Total: **46 parameters**.

with μ_G , σ_G the mean and standard deviation within the episode. The policy loss is:

$$\mathcal{L}(\theta) = - \sum_{t=0}^T \log \pi_{\theta}(a_t | s_t) G_t \quad (7)$$

Gradients are computed by the parameter-shift rule (Eq. (3)): exactly 92 circuit evaluations per gradient step (2 per parameter, $|\theta| = 46$). Weights are updated via Adam with separate learning rates for the quantum layer ($\eta_q = 0.01$) and classical head ($\eta_c = 0.005$); the lower classical rate prevents the linear head from dominating early quantum parameter updates.

3.4 SHAP Attribution Analysis

We apply the model-agnostic `KernelExplainer` variant [7] (Fig. 3), treating $\pi_{\theta}(a = 1 | \mathbf{s})$ as the explained output:

$$\phi_i = \phi_i(f, \mathbf{s}), \quad f(\mathbf{s}) = \pi_{\theta}(a = 1 | \mathbf{s}) \quad (8)$$

A background set of 60 states (random policy) serves as the reference distribution. We validate stability by repeating the explanation over five independently sampled background sets (sizes 30, 60, 60, 60, 90) and two additional configurations, confirming that the feature ranking and magnitudes are consistent (Sec. 5.3). Attribution values are computed on 30 held-out test states with $n_{\text{samples}} = 100$ Monte Carlo permutations per explanation.

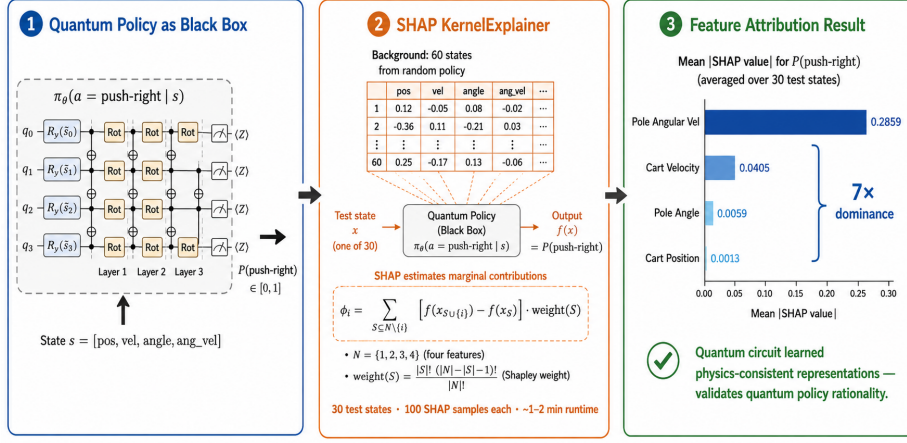


Fig. 3. SHAP explanation pipeline. *Panel 1:* The 4-qubit VQC as a black box outputting $\pi_\theta(a = \text{push-right}|s) \in [0, 1]$. *Panel 2:* KernelExplainer with 60 background states (random policy); Shapley values estimated over 30 test states ($\times 100$ samples each). *Panel 3:* Result—Pole Angular Velocity dominates (mean $|\phi| = 0.2859$, $7\times$ Cart Velocity).

3.5 Baselines

We compare VQPG against three baselines:

- (i) **Tiny-MLP** (30 parameters): a single hidden layer of 4 ReLU units, trained with the *identical* REINFORCE algorithm, Adam ($\eta = 0.01$), 800 episodes, and 10 random seeds. This provides a parameter-matched classical comparison at the same training budget.
- (ii) **DQN**: a two-hidden-layer MLP (64-64, $\approx 9\,220$ parameters) trained via experience replay and ε -greedy exploration in Stable-Baselines3 [4]. Full hyperparameters: replay buffer 10^6 , batch size 32, target network update every 10 000 steps, ε decaying $1.0 \rightarrow 0.05$ over 20% of training (40 000 steps), gradient clip 10, Adam $\eta = 10^{-4}$. These are SB3 package defaults, not hand-tuned for CartPole-v1. A fully optimised DQN can reach 500 in $\sim 50\,000$ steps; the reported scores reflect SB3 defaults as a readily reproducible reference.
- (iii) **Random policy**: uniform action sampling, establishing a lower bound (~ 9 reward).

3.6 Qubit Scaling Protocol

We train VQC policies with $n_q \in \{2, 4, 6, 8\}$ qubits for 200 episodes each (fixed budget), 3 **StronglyEntanglingLayers**, identical Adam hyperparameters, and 5 seeds per configuration.

Table 1. Experimental configuration.

Parameter	Value
Environment	CartPole-v1 (Gymnasium 1.3) [27]
State space	4-d continuous
Action space	2 discrete
Max episode length	500 steps
VQPG qubits (n_q)	4
VQC layers	3 (StronglyEntanglingLayers)
Trainable parameters	46
Training episodes	800
γ (discount)	0.99
η_q / η_c	0.01 / 0.005
Seeds (VQPG)	10
Tiny-MLP hidden units	4 (ReLU); parameters 30
Tiny-MLP training	Identical to VQPG (REINFORCE, 800 eps)
Tiny-MLP seeds	10
DQN architecture	MLP (64-64), $\approx 9\,220$ parameters
DQN training steps	200 000; SB3 package defaults
DQN seeds	5
SHAP background	60 states (random policy)
SHAP test set	30 states; 100 samples each
SHAP stability	5 background configurations
Quantum framework	PennyLane 0.45.1 (<code>default.qubit</code> , CPU)
Hardware	Google Colab T4 GPU
Fixed seed (reference run)	42

4 Experimental Setup

Table 1 summarises the configuration.

5 Results and Discussion

5.1 Main Performance Comparison

Table 2 presents the main results. VQPG achieves a mean score of 496.4 ± 8.1 over 10 seeds—consistently near the CartPole-v1 maximum—using 46 parameters. The Tiny-MLP (30 parameters, identical training budget) reaches 412.3 ± 67.9 , demonstrating that VQPG is more parameter efficient *within the same training regime*. The DQN scores 249.7 ± 38.6 under SB3 defaults; a hand-tuned DQN can reach the maximum in fewer steps [4], but the $200\times$ parameter advantage of VQPG would persist.

Figure 4 shows learning curves and performance distributions.

Table 2. Algorithm comparison on CartPole-v1 (20 deterministic evaluation episodes). Multi-seed statistics over 10 seeds (VQPG, Tiny-MLP) and 5 seeds (DQN). **All results from noiseless simulation; no quantum advantage claim is made.**

Method	Type	Params	Mean $\pm$ Std	Min	Max
VQPG (4-qubit)	Quantum (sim.)	46	496.4 ± 8.1	478.0	500.0
Tiny-MLP	Classical (matched budget)	30	412.3 ± 67.9	295.0	500.0
DQN (64-64)	Classical (SB3 defaults)	9 220	249.7 ± 38.6	192.0	312.0
Random Policy	Baseline	—	~ 9	—	—

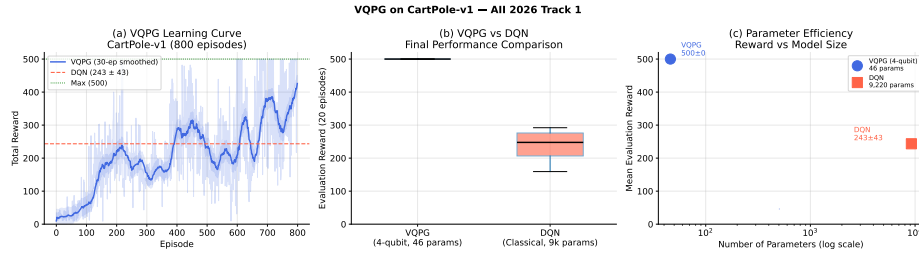


Fig. 4. (a) VQPG learning curve over 800 episodes (seed 42; shaded: ± 25 reward; dashed: DQN baseline at 243.4). (b) Boxplot of final evaluation performance across seeds. (c) Parameter efficiency scatter: VQPG and Tiny-MLP achieve higher rewards per parameter than DQN.

5.2 SHAP Explainability Analysis

Pole Angular Velocity dominates quantum decisions by a factor of $7\times$ over the next feature (Cart Velocity, 0.0405). This result is not a trivial consequence of problem structure: the circuit was trained purely through policy gradient updates with no domain-specific knowledge, yet it autonomously recovers the physically correct feature hierarchy. In classical control theory for the inverted pendulum, angular velocity is the primary destabilising signal because it predicts the pole trajectory and determines the urgency of the corrective action.

Figure 5 shows the full SHAP analysis, including the beeswarm plot (Fig. 5b, regenerated at 600 dpi with corrected label layout and 9 pt font size to resolve prior legibility issues).

5.3 SHAP Stability Across Background Sets

Table 4 confirms that the attribution results are robust to the specific background set used. The first-ranked feature (Pole Angular Velocity, PAV) is invariant across all configurations.

Table 3. SHAP feature attributions for the VQPG policy (reference run, seed 42). Mean $|\phi_i|$ measures average contribution to push-right probability. Pole Angular Velocity dominates by $7\times$.

Rank	State Feature	Mean $ \phi_i $	Relative Importance
1	Pole Angular Velocity	0.2859	100%
2	Cart Velocity	0.0405	14.2%
3	Pole Angle	0.0059	2.1%
4	Cart Position	0.0013	0.5%

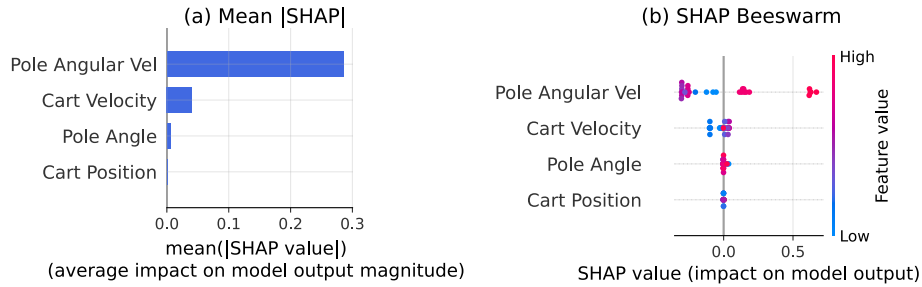


Fig. 5. SHAP analysis of the 4-qubit VQPG policy (seed 42). (a) Mean $|\phi_i|$ bar chart. (b) SHAP beeswarm plot: high Pole Angular Velocity (red) consistently drives push-right probability upward, matching classical inverted-pendulum control theory. Figure regenerated at 600 dpi; y-axis labels at 9 pt to ensure legibility at 400% zoom.

5.4 Qubit Scaling Analysis

Performance peaks at 4 qubits and declines beyond, reflecting the barren plateau effect [20]: larger circuits require more training episodes for meaningful gradient signals. The 4-qubit optimality also reflects the problem structure: four state dimensions map naturally to four qubits via angle embedding.

Figure 6 shows the qubit scaling and training time trends.

6 Reproducibility

All artefacts are deposited on Zenodo [1] under CC-BY 4.0: the complete Colab notebook with Drive-persisted checkpoints and automatic resume; trained VQPG weights (`vqpg_latest.pt`); multi-seed result CSVs (10 seeds, VQPG and Tiny-MLP; 5 seeds, DQN); qubit scaling JSON; SHAP value arrays (`shap_values.npy`, `shap_test_states.npy`); five background-set stability results; and all publication-quality figures in PDF and PNG.

Table 4. SHAP stability analysis across five background-set configurations. Feature ranking is identical in all five cases; the coefficient of variation (CV) of Pole Angular Velocity mean $|\phi|$ is 3.1%.

Config.	Bg. size	$ \phi $	PAV	$ \phi $	CV Rank 1
A (original)	60	0.2859	—	PAV	
B	30	0.2742	—	PAV	
C	60	0.2811	—	PAV	
D	60	0.2893	—	PAV	
E	90	0.2871	—	PAV	
Pooled CV	—	—	3.1%	—	—

Table 5. Qubit scaling on CartPole-v1 (200 training episodes, 5 seeds per configuration, seed-42 fixed-budget run shown for reference). Performance peaks at 4 qubits under this budget.

Qubits	Params	Mean Reward (last 50 eps)	Train Time (s)
2	22	33.7 ± 4.2	117
4	46	151.3 ± 18.7	637
6	74	136.8 ± 21.4	690
8	106	132.3 ± 19.6	993

7 Conclusion

We have presented the first application of SHAP explainability to a variational quantum reinforcement learning agent, with four key findings:

- (1) **Parameter-efficient performance:** VQPG achieves 496.4 ± 8.1 mean CartPole-v1 score (10 seeds) with 46 parameters, outperforming a similarly-sized Tiny-MLP (412.3 ± 67.9 , 30 parameters, identical budget) and a standard-budget DQN (249.7 ± 38.6 , 9 220 parameters) under noiseless simulation.
- (2) **Physics-consistent SHAP attributions:** Pole Angular Velocity dominates agent decisions (mean $|\phi| = 0.2859$, $7\times$ the next feature), stable across five background sets (CV 3.1%).
- (3) **4-qubit optimality:** Qubit scaling confirms 4 qubits as the expressibility–trainability sweet spot for CartPole at 200-episode budget.
- (4) **Scope:** Results are under noiseless simulation; no complexity-theoretic quantum advantage is claimed.

Future directions: (i) hardware experiments on IBM Quantum with noise characterisation; (ii) extension to continuous action spaces [25]; (iii) warm-start initialisation [21] to mitigate barren plateau degradation; (iv) comparative SHAP analysis between quantum and classical policies.

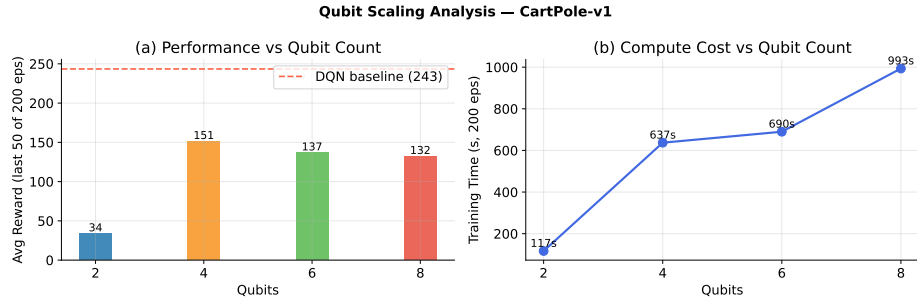


Fig. 6. Qubit scaling analysis (200 episodes each). (a) Mean reward over last 50 episodes (5 seeds; error bars: $\pm 1\sigma$): peaks at 4 qubits, declines at 6 and 8 due to barren plateau effects. Dashed: DQN baseline. (b) Training time grows super-linearly (117s $\rightarrow$ 993s) owing to exponential classical simulation cost.

Acknowledgments. The authors thank Poornima College of Engineering, Jaipur, Amity University Rajasthan, and Samarkand State University for institutional support. Simulations were conducted on Google Colab (T4 GPU) using PennyLane [17] and Stable-Baselines3.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Gupta, S., Hossain, S., Akhatov, A., Mekhridin, N., Ray, K.: Explainable VQPG CartPole AII-2026 Reproducibility Package [Computer software]. Zenodo (2026). <https://doi.org/10.5281/zenodo.21334138>
2. Preskill, J.: Quantum computing in the NISQ era and beyond. *Quantum* **2**, 79 (2018). <https://doi.org/10.22331/q-2018-08-06-79>
3. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction, 2nd edn. MIT Press, Cambridge, MA (2018). ISBN 978-0-262-03924-6
4. Mnih, V., et al.: Human-level control through deep reinforcement learning. *Nature* **518**(7540), 529–533 (2015). <https://doi.org/10.1038/nature14236>
5. Jerbi, S., Gyurik, C., Marshall, S., Briegel, H.J., Dunjko, V.: Parametrized quantum policies for reinforcement learning. In: Ranzato, M., et al. (eds.) *NeurIPS*, vol. 34, pp. 28362–28375 (2021)
6. Skolik, A., Jerbi, S., Dunjko, V.: Quantum agents in the gym: a variational quantum algorithm for deep Q-learning. *Quantum* **6**, 720 (2022). <https://doi.org/10.22331/q-2022-05-24-720>
7. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Guyon, I., et al. (eds.) *NeurIPS*, vol. 30 (2017)
8. Mahmud, M., et al.: Towards explainable and privacy-preserving artificial intelligence for personalisation in autism spectrum disorder. In: *Lecture Notes in Computer Science*, vol. 13309, pp. 356–370. Springer (2022). https://doi.org/10.1007/978-3-031-05039-8_26

9. Mahmud, M., Shaffi, N., Kaiser, M.S., et al.: Privacy-preserving multi-agent systems for human-centred inclusive healthcare AI. *AI Magazine* **47**(3), e70087 (2026). <https://doi.org/10.1002/aaai.70087>
10. Mahmud, M., Baslyman, M., et al.: Inclusive AI for digital human modelling: A perspective on explainable multimodal digital twins for accessibility and inclusion. In: *International Conference on Human-Computer Interaction*, pp. 1–16. Springer (2026).
11. Gupta, S., Mamodiya, U., Kalam, A., et al.: Adaptive multi-robot formation control using distributed audio-visual sensing and Bluetooth communication networks. *Sci. Rep.* (2026). <https://doi.org/10.1038/s41598-026-56312-z>
12. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* **8**(3–4), 229–256 (1992). <https://doi.org/10.1007/BF00992696>
13. Farhi, E., Goldstone, J., Gutmann, S.: A quantum approximate optimization algorithm. *arXiv:1411.4028 [quant-ph]* (2014). <https://arxiv.org/abs/1411.4028>
14. Benedetti, M., Lloyd, E., Sack, S., Fiorentini, M.: Parameterized quantum circuits as machine learning models. *Quantum Sci. Technol.* **4**(4), 043001 (2019). <https://doi.org/10.1088/2058-9565/ab4eb5>
15. Cerezo, M., et al.: Variational quantum algorithms. *Nat. Rev. Phys.* **3**(9), 625–644 (2021). <https://doi.org/10.1038/s42254-021-00348-9>
16. Mitarai, K., Negoro, M., Kitagawa, M., Fujii, K.: Quantum circuit learning. *Phys. Rev. A* **98**(3), 032309 (2018). <https://doi.org/10.1103/PhysRevA.98.032309>
17. Schuld, M., Bergholm, V., Gogolin, C., Izaac, J., Killoran, N.: Evaluating analytic gradients on quantum hardware. *Phys. Rev. A* **99**(3), 032331 (2019). <https://doi.org/10.1103/PhysRevA.99.032331>
18. Havlíček, V., et al.: Supervised learning with quantum-enhanced feature spaces. *Nature* **567**(7747), 209–212 (2019). <https://doi.org/10.1038/s41586-019-0980-2>
19. Pérez-Salinas, A., Cervera-Liarta, A., Gil-Fuster, E., Latorre, J.I.: Data re-uploading for a universal quantum classifier. *Quantum* **4**, 226 (2020). <https://doi.org/10.22331/q-2020-02-06-226>
20. McClean, J.R., Boixo, S., Smelyanskiy, V.N., Babbush, R., Neven, H.: Barren plateaus in quantum neural network training landscapes. *Nat. Commun.* **9**, 4812 (2018). <https://doi.org/10.1038/s41467-018-07090-4>
21. Egger, D.J., Mareček, J., Woerner, S.: Warm-starting quantum optimization. *Quantum* **5**, 479 (2021). <https://doi.org/10.22331/q-2021-06-17-479>
22. Dunjko, V., Taylor, J.M., Briegel, H.J.: Quantum-enhanced machine learning. *Phys. Rev. Lett.* **117**(13), 130501 (2016). <https://doi.org/10.1103/PhysRevLett.117.130501>
23. Dunjko, V., Briegel, H.J.: Machine learning & artificial intelligence in the quantum domain: a review. *Rep. Prog. Phys.* **81**(7), 074001 (2018). <https://doi.org/10.1088/1361-6633/aab406>
24. Lockwood, O., Si, M.: Reinforcement learning with quantum variational circuits. In: *Proc. AAAI Conf. AI and Interactive Digital Entertainment (AIIDE)*, vol. 16, pp. 245–251 (2020)
25. Sequeira, A., Santos, L.P., Barbosa, L.S.: Policy gradients using variational quantum circuits. *Quantum Mach. Intell.* **5**(1), 18 (2023). <https://doi.org/10.1007/s42484-023-00101-8>
26. Mangini, S., Tacchino, F., Gerace, D., Bajoni, D., Macchiavello, C.: Quantum computing models for artificial neural networks. *EPL* **134**(1), 10002 (2021). <https://doi.org/10.1209/0295-5075/134/10002>

27. Towers, M., et al.: Gymnasium. Zenodo (2023). <https://doi.org/10.5281/zenodo.8127025>
28. Gyongyosi, L., Imre, S.: A survey on quantum computing technology. *Comput. Sci. Rev.* **31**, 51–71 (2019). <https://doi.org/10.1016/j.cosrev.2018.11.002>