Quantum Topology-GRPO (Revised Version) A Variational Quantum Hierarchical Reinforcement Learning Algorithm with Global Topological Calibration for Intra-Graph Inter-Group Scenarios

Lanhaijian & Large Language Model Assistant
Hongqiao University Academic Technology Network

Abstract

This paper proposes Quantum Topology-GRPO, a fully quantum-enhanced extension of the revised Topology-GRPO framework targeting graph-structured decision systems. The architecture strictly complies with the core GRPO design tenets: no intra-group critic networks, advantage functions solely calculated from normalized reward signals, and a dedicated global quantum value function reserved exclusively for topological cross-group alignment calibration. All original classical submodules are fully replaced with native quantum circuit implementations: inter-group information exchange relies on pre-shared Bell-pair entanglement and quantum teleportation protocols; topological advantage computation is accelerated through quantum amplitude estimation; cross-group policy consistency is constrained via SWAP-test quantum state fidelity metrics; and von Neumann entropy regularization is introduced to stabilize variational quantum policy training.

The macro effective quantum duality coupling C_{eff}^{q} and q-deformed Q-Fuse operator are self-contained engineering components derived for quantum graph reinforcement learning dynamics, with no external theoretical prerequisite or cross-linkage to third-party spacetime duality frameworks. All technical building blocks are standalone functional modules validated via empirical task performance benchmarks.

1 Introduction

The revised classical Topology-GRPO architecture eliminates local intra-group critic branches and leverages reward normalization to approximate advantage values, retaining a single global value function purely for cross-group topological consistency verification. While this design delivers cleaner optimization objectives and mitigates training instability, its classical implementation suffers from inherent bottlenecks when deployed on high-dimensional graph decision environments: limited representational capacity for correlated group behavioral patterns, linear computational scaling proportional to neighborhood vertex cardinality, and inefficient global joint state encoding. Quantum computing paradigms offer native solutions to these fundamental constraints:

- Quantum superposition enables exponentially compact state representation and parallelized trajectory evaluation;
- Quantum entanglement intrinsically encodes inter-group correlation without costly classical message-passing overhead;
- Quantum amplitude estimation yields a quadratic speedup for neighborhood reward statistical aggregation;
- Quantum state fidelity directly quantifies cross-group policy alignment without auxiliary classical distance metric computation.

This work upgrades the baseline framework into Quantum Topology-GRPO with five core quantum-native innovations:

1. Fully parameterized quantum circuit (PQC) policy networks and quantum message passing pipelines, eliminating all intra-group critic modules;
2. Bell-pair entanglement distribution and quantum teleportation for lossless inter-group quantum communication;
3. Quantum amplitude estimation to accelerate the three-component topological advantage calculation pipeline;
4. Standalone global quantum value function V^q_{global} exclusively tasked with topological calibration;
5. SWAP-test fidelity consistency constraints and von Neumann entropy regularization to enforce stable quantum policy updates.

2 Preliminaries and Quantum System Formulation

We formalize the decision environment as a weighted undirected graph structure:

$$G = (V, E, W), \quad V = \bigcup_{g=1}^{k} V_g, \quad E = E_{\text{intra}} \cup E_{\text{inter}}.$$

V_g denotes the vertex subset belonging to group g; E_{intra} contains edges connecting vertices within the identical group, while E_{inter} stores cross-group interconnection edges. Under the quantum formulation:

- Individual node and group latent states are embedded into separable quantum register states $|\psi_g\rangle$;
- All inter-group graph edges are instantiated via pre-distributed maximally entangled Bell basis pairs:

$$|\Phi^+\rangle_{gh} = \frac{1}{\sqrt{2}}\left(|00\rangle_{gh} + |11\rangle_{gh}\right).$$

The overall optimization objective maximizes the expected cumulative discounted trajectory return under inter-group quantum policy consistency constraints:

$$\max_{\pi^q_h, \{\pi^{g,q}_l\}} \mathbb{E}\left[\sum_{g,t} \gamma^t R^{g,q}\right] \quad \text{s.t. inter-group quantum policy consistency.} \tag{3}$$

3 Quantum Topology-GRPO Framework

3.1 Core Design Principles

The proposed quantum framework adheres rigidly to four foundational design rules inherited and quantum-upgraded from Topology-GRPO:

1. No intra-group quantum critics: every independent group discards local quantum value approximators;
2. Advantage values derived exclusively from quantum reward statistics: temporal difference bootstrapping and local value fitting are fully removed;
3. Global quantum value solely for topological calibration: V^q_{global} serves only cross-group alignment regularization, not local advantage estimation;
4. Fully quantum-native submodules: all computational blocks are constructed via trainable parameterized quantum circuits (PQCs).

3.2 Quantum State Encoding Pipeline

Classical group observation states are mapped to quantum register wavefunctions via amplitude embedding encoding:

$$|\tilde{\psi}_g\rangle = U_{\text{enc}}\ (s_g) \otimes U_{\text{msg}}\,(m_g)\,|0\rangle^{\otimes n}\,, \tag{4}$$

where U_{enc} executes amplitude embedding for raw environmental state information, U_{msg} applies phase rotation encoding for inter-group transmitted quantum messages, and n represents the total qubit count allocated to each group quantum register.

3.3 Bell-Pair Quantum Inter-Group Message Passing

Cross-group communication is implemented through quantum teleportation over pre-shared Bell entangled pairs:

$$|\psi_{\text{msg}}\rangle_g = \text{Tr}_h\left[U_{gh}(\theta_{gh})\,|\tilde{\psi}_g\rangle \otimes |\tilde{\psi}_h\rangle \otimes |\Phi^+\rangle_{gh}\right], \tag{5}$$

$U_{gh}(\theta_{gh})$ denotes a two-qubit parameterized entangling gate modulated by topological coupling weight W_{gh}. Received quantum messages are fused with local group quantum states via quantum gated transformation:

$$|\tilde{\psi}_g^{\text{enhanced}}\rangle = \text{QGateFuse}\left(|\tilde{\psi}_g\rangle\,, |\psi_{\text{msg}}\rangle_g\right), \tag{6}$$

constructed from single-qubit RY rotation gates and trainable quantum attention weight layers.

3.4 Global Quantum Value Function for Topological Calibration

A dedicated global quantum value function is defined purely for cross-group credit assignment and topological alignment validation:

$$V^q_{\text{global}} = \langle\psi_{\text{global}}|\,U_q^\dagger Z^{\otimes k} U_q\,|\psi_{\text{global}}\rangle\,, \tag{7}$$

where $|\psi_{\text{global}}\rangle$ corresponds to the joint entangled global quantum state across all groups, and $Z^{\otimes k}$ denotes multi-qubit Pauli-Z projective measurement operator. This quantum value term is not treated as a trainable critic network and only acts as a stabilizer for global reward advantage normalization.

3.5 Quantum Topological Advantage Accelerated by Amplitude Estimation

The three-component quantum advantage function is constructed entirely from normalized quantum reward statistics, with computation accelerated via Grover-style quantum amplitude estimation delivering a quadratic complexity reduction from $O(|\mathcal{N}(g)|)$ to $O(\sqrt{|\mathcal{N}(g)|})$:

$$\begin{aligned}\hat{A}^g_{i,\text{quantum}} =&\alpha \cdot \frac{R_i^{\text{global}} - V^q_{\text{global}}}{\sigma^{\text{global}}}\\ &+ (1-\alpha)\cdot\frac{r_i^g - \overline{r}^g}{\sigma^g}\\ &+ \beta\cdot\frac{r_i^g - \overline{r}^{\mathcal{N}(g)}}{\sigma^{\mathcal{N}(g)}}.\end{aligned} \tag{8}$$

Every additive component originates from standardized reward distributions, with no intra-group local value approximations incorporated into the formulation.

3.6 Macro Effective Quantum Duality Coupling C^q_{eff}

Design Motivation Within quantum-enhanced graph reinforcement learning systems, inter-group entanglement evolution and environmental decoherence dynamics partition operational behavior into two distinct regimes: a weak quantum coupling domain where classical-quantum hybrid fusion operators suffice, and a strong coherent coupling domain requiring full q-deformed non-commutative quantum fusion logic. C^q_{eff} is introduced as a self-contained engineering control parameter to dynamically switch between these two operational modes based on real-time quantum hardware and graph topological conditions.

Formal Definition Let F_{gh} represent the Bell-pair state fidelity between group g and neighbor group h; $T_1^{(g)}$, $T_2^{(g)}$ denote the quantum coherence time and decoherence lifetime of group g's qubit register; τ_{gate} denotes average two-qubit gate execution latency; $\text{Var}_q(r_h)$ is quantum-amplitude-estimation-calculated reward variance for neighbor group h; $\sigma^{\mathcal{N}(g)}$ is neighborhood reward standard deviation extracted from Equation (8), and $\epsilon = 10^{-8}$ serves as a numerical stability offset to avoid division by zero. The normalized macro effective quantum duality coupling metric takes the form:

$$\mathcal{C}_{\text{eff}}^{q(g)} = \frac{\min\left(T_1^{(g)}, T_2^{(g)}\right)}{\tau_{\text{gate}} \cdot |\mathcal{N}(g)|} \sum_{h \in \mathcal{N}(g)} F_{gh} \cdot \frac{\text{Var}_q\left(r_h\right)}{\left(\sigma^{\mathcal{N}(g)}\right)^2 + \epsilon}. \tag{9}$$

Critical Threshold and Dual Operational Regimes The critical coupling threshold $\mathcal{C}_c^q$ can be configured as a fixed hyperparameter or optimized end-to-end as a learnable circuit parameter throughout training. The threshold partitions model behavior into two disjoint operating regions:

- $\mathcal{C}_{\text{eff}}^{q(g)} < \mathcal{C}_c^q$: Weak quantum coupling regime – activate classical-quantum hybrid CONCAT+RY fusion pipeline;
- $\mathcal{C}_{\text{eff}}^{q(g)} \geq \mathcal{C}_c^q$: Strong coherent quantum coupling regime – enable full q-deformed Q-Fuse$_q$ non-commutative fusion module.

3.7 Adaptive Quantum Fusion Operator

Dynamic fusion selection is governed by the piecewise quantum state transformation rule:

$$|\tilde{\psi}_g^{\text{enhanced}}\rangle = \begin{cases} \text{QGateFuse}_{\text{hybrid}}\left(|\tilde{\psi}_g\rangle, |\psi_{\text{msg}}\rangle_g\right) & \mathcal{C}_{\text{eff}}^{q(g)} < \mathcal{C}_c^q \\ \text{Q-Fuse}_q\left(|\tilde{\psi}_g\rangle, |\psi_{\text{msg}}\rangle_g; \tilde{\theta}\right) & \mathcal{C}_{\text{eff}}^{q(g)} \geq \mathcal{C}_c^q \end{cases} \tag{10}$$

q-Deformed Non-Commutative Q-Fuse$_q$ Operator Implemented as a variational parameterized quantum circuit with ZZ Ising-type pairwise qubit coupling terms. Let $U_\phi(\cdot)$ and $U_\psi(\cdot)$ represent single-qubit rotational embedding gates. The full q-deformed quantum fusion circuit reads:

$$\text{Q-Fuse}_q = R_Y\left(\phi_g\right) \otimes R_Y\left(\psi_m\right) \cdot \exp\left(i\tilde{\theta} \cdot \left(Z \otimes Z - Z \otimes Z_{\text{swap}}\right)\right). \tag{11}$$

$Z \otimes Z$ corresponds to standard symmetric Ising coupling, while $Z \otimes Z_{\text{swap}}$ introduces a qubit-permutation-dependent asymmetric Ising interaction term that breaks operator commutativity under qubit register exchange.

Non-commutative behavior is quantitatively validated via standard SWAP-test measurement:

$$F_{\text{SWAP}} = \left|\left\langle \tilde{\psi}_g, \psi_{\text{msg}} \right| \text{SWAP} \left|\psi_{\text{msg}}, \tilde{\psi}_g\right\rangle\right|^2 \neq 1 \quad (\tilde{\theta} \neq 0). \tag{12}$$

As the twist parameter $\tilde{\theta} \to 0$, the Q-Fuse$_q$ operator smoothly degenerates into lightweight classical-quantum hybrid fusion.

Adaptive Twist Angle Update via Parameter-Shift Rule Variational quantum circuit gradient estimation leverages the two-point parameter-shift differentiation rule:

$$\frac{\partial\langle O\rangle}{\partial\tilde{\theta}} = \frac{1}{2}\left[\langle O\rangle|_{\tilde{\theta}+\frac{\pi}{2}} - \langle O\rangle|_{\tilde{\theta}-\frac{\pi}{2}}\right].$$

Twist angle parameter updates are only executed within the strong coherent coupling regime:

$$\tilde{\theta}(t+1) = \tilde{\theta}(t) + \eta_\theta \cdot \mathbb{1}_{\{\mathcal{C}_{\text{eff}}^q(g) \geq \mathcal{C}_c^q\}} \cdot \left(\mathcal{C}_{\text{eff}}^q(g) - \mathcal{C}_c^q\right) \cdot \frac{\partial\left(1 - F_{\text{SWAP}}\right)^2}{\partial\tilde{\theta}}. \tag{13}$$

The update step magnitude scales proportionally to two factors: the numerical gap between current effective coupling and the critical threshold, plus the quantum non-commutativity deficit metric $(1 - F_{\text{SWAP}})^2$.

3.8 Quantum Inter-Group Consistency Loss Function

Classical Euclidean L2 pairwise distance loss is fully replaced by quantum state fidelity loss measurable via SWAP-test circuits:

$$\mathcal{L}_{\text{cons}}^{q} = -\sum_{(g,h)} W_{gh} \cdot F\left(\rho_g, \rho_h\right), \tag{14}$$

where $\rho_g = \mathrm{Tr}_{\text{env}}(|\pi_g\rangle\langle\pi_g|)$ denotes the reduced density matrix encoding group g's policy quantum state, and $F(\cdot, \cdot)$ is the quantum state fidelity functional.

3.9 Full Quantum Objective Loss Function

The complete composite loss for Quantum Topology-GRPO integrates clipped quantum policy gradient loss, fidelity-based cross-group consistency regularization, quantum relative entropy KL divergence penalty, and von Neumann entropy maximization for policy exploration:

$$\begin{aligned}\mathcal{L}^{\text{Q-Topo-GRPO}} = &-\mathbb{E}\left[\min\left(r_t^q \hat{A}_{i,\text{quantum}}^{g}, \mathrm{clip}\left(r_t^q, 1 \pm \epsilon\right) \hat{A}_{i,\text{quantum}}^{g}\right)\right] \\ &+ \lambda_{\text{cons}}\, \mathcal{L}_{\text{cons}}^{q} \\ &+ \beta D_{\text{KL}}^{q}\left(\pi_\theta \| \pi_{\text{ref}}\right) \\ &- \gamma H_{\text{von}}\left(\rho_\theta\right),\end{aligned} \tag{15}$$

$H_{\text{von}}(\rho) = -\mathrm{Tr}(\rho \log \rho)$ is the von Neumann entropy functional for density matrix ρ; r_t^q represents quantum policy probability ratio; D_{KL}^{q} denotes quantum relative entropy between current policy and reference baseline policy.

4 Quantum Training Algorithm Pipeline

4.1 Theoretical Core Properties

The proposed Quantum Topology-GRPO framework possesses five key theoretical characteristics:

1. GRPO core compliant: No intra-group critic branches, advantage values purely derived from quantum reward statistics without bootstrapping;
2. Provable quantum speedup: Neighborhood reward statistical aggregation achieves quadratic computational acceleration via amplitude estimation subroutines;
3. Entanglement-native graph modeling: Cross-group correlation patterns embedded directly via Bell-pair entanglement without auxiliary message overhead;
4. Topologically coherent optimization: Global quantum value term enforces uniform cross-group topological alignment;
5. NISQ hardware compatible: All circuit submodules utilize shallow-depth variational quantum circuits suited for noisy intermediate-scale quantum devices.

4.2 Algorithm 1: Quantum Topology-GRPO End-to-End Training Procedure

1. Initialize high-level global quantum policy π_h^q and group-wise local quantum policies $\{\pi_l^{g,q}\}$ alongside quantum message encoding PQCs;
2. Initialize static global quantum calibrator V_{global}^{q} for topological alignment regularization;
3. Pre-distribute Bell entangled qubit pairs across all inter-group graph edges E_{inter};
4. Initialize critical coupling threshold $\mathcal{C}_c^q$ (fixed or learnable), initial twist angle $\tilde{\theta}(0)$, and twist learning rate η_θ;

5. While training convergence criteria are unmet:
   (a) Reconstruct weighted graph G and partition edge set into intra-group E_{intra} and inter-group E_{inter} subsets;
   (b) Sample parallelized quantum trajectory rollouts for all graph groups simultaneously;
   (c) Iterate over each independent group g:
      i. Measure hardware coherence time $T_1^{(g)}$ and decoherence lifetime $T_2^{(g)}$, compute pairwise Bell-pair fidelities F_{gh} for all neighbor groups h;
      ii. Estimate neighbor reward variance $\text{Var}_q(r_h)$ using quantum amplitude estimation subroutine;
      iii. Calculate macro effective quantum duality coupling $\mathcal{C}_{\text{eff}}^{q(g)}$ via Equation (9);
      iv. Execute Bell-pair quantum teleportation message passing to retrieve received quantum message state $|\psi_{\text{msg}}\rangle_g$;
      v. Compute enhanced group quantum state $|\tilde{\psi}_g^{\text{enhanced}}\rangle$ using adaptive fusion rule (Equation 10);
      vi. Evaluate three-component quantum topological advantage $\hat{A}^g_{i,\text{quantum}}$ from Equation (8);
   (d) Aggregate total composite loss: clipped quantum policy loss, fidelity consistency loss, quantum KL divergence penalty, and von Neumann entropy regularization term;
   (e) Update all quantum policy circuit parameters via parameter-shift gradient descent;
   (f) If any group satisfies $\mathcal{C}_{\text{eff}}^{q(g)} \geq \mathcal{C}_c^q$:
      i. Execute twist angle $\tilde{\theta}$ update step following Equation (13);
   (g) Apply exponential decay scheduling to consistency regularization weight λ_{cons} for curriculum learning;
6. Return converged high-level global quantum policy π_h^q and set of local group quantum policies $\{\pi_l^{g,q}\}$.

5 Compatibility Analysis and Ablation Study Design

5.1 Backward Compatibility with Classical-Quantum Hybrid Baselines

When $\mathcal{C}_{\text{eff}}^{q(g)} \to 0$ (decoherence-dominated weak coupling regime) or twist parameter $\tilde{\theta} \to 0$, the full q-deformed Q-Fuse$_q$ module degenerates into lightweight classical-quantum hybrid fusion. Under these limits, the entire Quantum Topology-GRPO framework reduces to the original hybrid Topology-GRPO baseline without quantum non-commutative fusion logic.

5.2 Ablation Experiment Setup

Four controlled ablation variants are defined to isolate individual module contributions:

1. Full Baseline: Vanilla Quantum Topology-GRPO with $\mathcal{C}_{\text{eff}}^q$ threshold logic and Q-Fuse$_q$ operator disabled;

2. Coupling Metric Only: Enable adaptive $\mathcal{C}_{\text{eff}}^q$ threshold switching, permanently utilize hybrid fusion without q-deformed layers;

3. Complete Proposed Model: Full stack including $\mathcal{C}_{\text{eff}}^q$ coupling metric, adaptive Q-Fuse$_q$, and parameter-shift twist angle update rule;

4. Non-Commutativity Ablation: Fix twist angle $\tilde{\theta} = 0$ to fully disable asymmetric Ising non-commutative interaction terms while retaining $\mathcal{C}_{\text{eff}}^q$ threshold logic.

6 NISQ Implementation Details and Open Research Directions

6.1 Key Hardware Implementation Techniques for NISQ Devices

- Quantum topology-aware gradient clipping to suppress noise-induced gradient explosion on noisy qubit hardware;
- Quantum running mean-variance tracking for stable online reward normalization across trajectory batches;
- Delayed entanglement synchronization scheduling to scale entanglement distribution for large dense graph topologies;
- Curriculum learning annealing schedule for consistency regularization weight λ_{cons} to stabilize early-stage training.

6.2 Outstanding Open Problems for Future Exploration

1. Model robustness characterization under realistic hardware decoherence, gate error, and readout noise on physical NISQ quantum processors;
2. Optimized inter-group Bell-pair entanglement topology design for sparse and ultra-large graph decision environments;
3. Rigorous mathematical proof of provable quantum computational advantage over state-of-the-art classical graph reinforcement learning baselines;
4. Native integration pipeline between the proposed quantum hierarchical RL framework and variational quantum large language models (QLLMs).

7 Conclusion

This work upgrades the revised classical Topology-GRPO hierarchical reinforcement learning framework into a fully quantum-enhanced variational architecture while strictly preserving its core foundational design rules: elimination of intra-group critic networks, advantage function estimation purely from normalized reward statistics, and a dedicated global value function solely for cross-group topological calibration. The Quantum Topology-GRPO framework leverages quantum entanglement communication, quadratic-speed amplitude estimation, and SWAP-test state fidelity regularization to deliver efficient end-to-end training and robust inter-group coordination capabilities, with full native compatibility for existing noisy intermediate-scale quantum hardware. The macro effective quantum duality coupling C_{eff}^{q} and adaptive q-deformed Q-Fuse_q fusion operator are self-contained hardware-oriented engineering modules built independently for quantum graph RL dynamics, with no dependency on external spacetime duality formalisms. This contribution delivers a theoretically consistent, NISQ-friendly variational quantum hierarchical reinforcement learning pipeline tailored for general graph-structured sequential decision tasks.