

PPO-Based Quantum-Aware Resource Optimization for Decoy-State BB84 in FSO Wireless Networks

Prerna Chaudhary, Nikita Kumari and Manav R. Bhatnagar

Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, India

Emails: {prerna.chaudhary, eez248400, manav}@ee.iitd.ac.in

Abstract—Hybrid quantum-classical wireless networks are a promising paradigm for secure next-generation systems, yet practical deployment is challenged by dynamic resource allocation, atmospheric turbulence, and cross-domain noise. We propose a quantum-aware deep reinforcement learning framework for adaptive optimization of decoy-state BB84 communication over free-space optical (FSO) channels. Our proximal policy optimization (PPO) agent incorporates quantum bit error rate (QBER) and Rytov variance into its state representation for intelligent cross-layer decision-making. To mitigate action lag under rapid fading, an offline-trained supervised learning model proactively forecasts turbulence impairments. Evaluated against Helstrom bounds and Gamma-Gamma fading models, simulations demonstrate our framework achieves significantly higher secret key rates, reduced QBER, and enhanced effective throughput compared to continuous-action DRL baselines and static heuristics.

Index Terms—quantum key distribution, machine learning, deep reinforcement learning, key exchange rate, bit error rate

I. INTRODUCTION

The rapid evolution of wireless communication systems has created an urgent demand for enhanced security, capacity, and resilience. As networks approach their theoretical limits, quantum technologies have emerged as a promising solution. Quantum-assisted wireless communication, which integrates quantum key distribution (QKD) techniques, offers fundamentally secure communication and improved robustness against adversarial attacks and channel impairments [1].

Despite these advantages, the integration of quantum and classical communication remains highly challenging. The hybrid nature of such systems introduces complexities in resource allocation, noise management, and protocol coordination [2]. Efficiently distributing limited quantum resources, mitigating impairments in both quantum and classical channels, and designing seamless hybrid protocols are essential to unlocking the potential of quantum-assisted wireless networks [3].

Machine learning (ML), and in particular deep reinforcement learning (DRL), has demonstrated strong capabilities in optimizing large-scale, dynamic systems [4]–[6]. In the context of quantum-assisted wireless communication, ML enables adaptive resource management, predictive noise mitigation, and cross-domain optimization, overcoming the limitations of static heuristics and rule-based approaches [7].

The potential applications of quantum-assisted wireless communication are wide-ranging. Use cases include secure 6G

networks, where quantum layers provide advanced protection for data transmission; quantum-secured internet of things (IoT), enabling resilient device-level communication in smart environments; and quantum smart grids, where quantum-encrypted control signals safeguard energy infrastructure against cyber threats [8]–[12]. Real-world demonstrations, such as hybrid quantum-classical QKD implementations and QKD over free-space optical (FSO) channels, further highlight the feasibility of deploying such systems in wireless environments [9], [13]–[15].

In this paper, we propose an ML driven optimization framework for quantum-assisted wireless networks. Unlike conventional resource allocation schemes that rely on heuristics or fixed policies, our approach leverages DRL for dynamic allocation of quantum and classical resources, while incorporating supervised regression models for proactive noise prediction and mitigation. Through extensive simulations based on realistic hybrid channel characteristics, we demonstrate significant improvements in key exchange rate (KER), quantum bit error rate (QBER), and system capacity relative to baseline methods. These results underline the transformative role of ML in enabling secure, efficient, and adaptive quantum-assisted wireless communication for next-generation networks.

The main contributions of this paper are as follows:

- 1) **Physics-Informed DRL:** A model guided PPO architecture integrates analytical Gamma-Gamma fading directly into the action space. A deterministic safety layer leverages QBER monotonicity to guarantee constraint adherence, overcoming traditional black-box limits.
- 2) **Proactive Forecasting:** To prevent reactive action lag, an embedded supervised multi-layer perceptron (MLP) forecasts rapid atmospheric turbulence (σ_R^2), enabling proactive reallocation of hybrid resources ahead of channel degradation.
- 3) **Cross-Layer Quantum Awareness:** A novel state representation couples continuous quantum metrology (Rytov variance, QBER) with classical FSO states, allowing the agent to dynamically co-schedule time, power, and pulse intensities.
- 4) **Analytical Validation:** Simulations grounded in Gamma-Gamma fading demonstrate significant improvements in KER, QBER, and effective throughput over static heuristics and continuous-action DRL baselines.

II. SYSTEM MODEL AND METHODOLOGY

We study a hybrid quantum-classical wireless link that co-schedules (i) a quantum optical channel for key distribution and (ii) a classical FSO channel for data reconciliation, and control. The integration of quantum and classical layers requires a unified control plane capable of managing diverse physical constraints in real time. In this framework, the central controller acts as a decision-making agent that observes instantaneous channel states, including Rytov variance and classical signal-to-noise ratio (SNR), to determine the optimal allocation of transmission resources. By dynamically adjusting the split between QKD and classical data transmission, the system maximizes secure communication utility while strictly adhering to QBER and outage probability thresholds. Time is slotted into intervals of duration T_s , during which the controller splits time and power between the two planes. To ensure tractability, we assume negligible clock drift and stationary background noise within each slot. Additionally, the mapping from Rytov variance to Gamma-Gamma parameters is pre-calibrated, and standard finite-key penalties can be seamlessly incorporated into the security constraints if required.

A. Network Timing, Power, and Decision Variables

In each slot of duration T_s , the quantum transmitter operates at f_{clk} pulses/s, yielding $N_q(t) = \tau_t f_{\text{clk}} T_s$ pulses, where $\tau_t \in [0, 1]$ is the quantum time fraction. The total transmit power is split between the quantum source (P_t^q) and classical FSO laser (P_t^c), satisfying $P_t^q + P_t^c \leq P_{\text{max}}$. The controller optimizes the action vector:

$$a_t = \{\tau_t, P_t^q, P_t^c, \mu_t, R_t^c, f_t\}, \quad (1)$$

where $\mu_t \in \mathcal{M}$ is the mean photon number (including optional decoy states), $R_t^c \in \mathcal{R}$ is the classical coding rate, and $f_t \geq 1$ is the reconciliation efficiency.

B. Quantum Optical Channel

The quantum plane uses a FSO path with turbulence and background noise.

1) *Turbulence and scintillation*: Let $I > 0$ denote the normalized received intensity. Under weak-to-strong turbulence, I is modeled by the Gamma-Gamma turbulence model [14].

$$f_I(i) = \frac{2(\alpha\beta)^{\frac{\alpha+\beta}{2}}}{\Gamma(\alpha)\Gamma(\beta)} i^{\frac{\alpha+\beta}{2}-1} K_{\alpha-\beta}(2\sqrt{\alpha\beta}i), \quad i > 0, \quad (2)$$

with shape parameters (α, β) determined by propagation geometry and the Rytov variance σ_R^2 . For plane waves, $\sigma_R^2 = 1.23 C_n^2 k^{7/6} L^{11/6}$ with $k = \frac{2\pi}{\lambda}$. A monotone mapping $\sigma_R^2 \mapsto (\alpha, \beta)$ is assumed.

2) *Photon detection model*: Per pulse with intensity μ , detection efficiency η , and instantaneous intensity I , the signal click probability is

$$p_{\text{sig}}(I) = 1 - \exp(-\eta\mu I). \quad (3)$$

Let p_d be the per-gate dark/background click probability and e_d the misalignment error. The slot-wise gain (click probability) and instantaneous QBER are

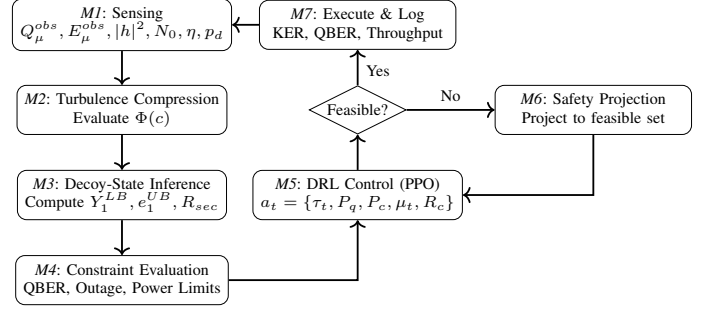


Fig. 1. Per-slot ML-driven optimization workflow for the hybrid quantum-classical link.

$$Q_\mu(I) \approx p_{\text{sig}}(I) + 2p_d, \quad (4)$$

$$E_\mu(I) = \frac{e_d p_{\text{sig}}(I) + p_d}{Q_\mu(I)}. \quad (5)$$

3) *Averaging over turbulence*: Define the Laplace transform of I ,

$$\Phi(c) := \mathbb{E}[e^{-cI}] = \sum_{k=0}^{\infty} \frac{(-c)^k}{k!} \frac{\Gamma(\alpha+k)\Gamma(\beta+k)}{\Gamma(\alpha)\Gamma(\beta)} (\alpha\beta)^{-k}, \quad (6)$$

which converges rapidly and is numerically stable. With $c = \eta\mu, c > 0$,

$$\bar{Q}_\mu = \mathbb{E}[Q_\mu(I)] = 1 - \Phi(c) + 2p_d, \quad (7)$$

$$\bar{E}_\mu = \frac{e_d [1 - \Phi(c)] + p_d}{1 - \Phi(c) + 2p_d}. \quad (8)$$

Monotonicity: Since $\Phi'(c) = -\mathbb{E}[Ie^{-cI}] < 0$, $(1 - \Phi(c))$ increases with c , which enables safe one-dimensional bisection for constraints on $\bar{E}_\mu$.

4) *Decoy-state estimation and secret key rate*: Let μ_s be the signal intensity and μ_w a weak decoy. Observables per intensity are (Q_μ, E_μ) . With vacuum yield $Y_0 \approx 2p_d$ and $e_0 = \frac{1}{2}$,

$$Y_1^{\text{LB}} = \frac{\mu_s}{\mu_s \mu_w - \mu_w^2} \left(Q_{\mu_w} e^{\mu_w} - \frac{\mu_w^2}{\mu_s^2} Q_{\mu_s} e^{\mu_s} - \frac{\mu_s^2 - \mu_w^2}{\mu_s^2} Y_0 \right), \quad (9)$$

$$e_1^{\text{UB}} = \frac{E_{\mu_w} Q_{\mu_w} e^{\mu_w} - e_0 Y_0}{\mu_w Y_1^{\text{LB}}}. \quad (10)$$

The secret key fraction per pulse is

$$R_{\text{sec}} \geq q \left[-Q_{\mu_s} f H_2(E_{\mu_s}) + Y_1^{\text{LB}} \mu_s e^{-\mu_s} (1 - H_2(e_1^{\text{UB}})) \right], \quad (11)$$

with sifting $q = \frac{1}{2}$ and binary entropy $H_2(\cdot)$. The key exchange rate in slot t is

$$\text{KER}_t = \tau_t f_{\text{clk}} R_{\text{sec},t}. \quad (12)$$

C. Classical Wireless Channel

The classical plane carries data and QKD reconciliation.

1) *Baseband model, fading, and SNR*: The received signal is $y = hx + n$, with $n \sim \mathcal{CN}(0, N_0)$ and bandwidth W . The instantaneous SNR is

$$\gamma_t = \frac{P_t^c |h_t|^2}{N_0 W}. \quad (13)$$

We consider Rayleigh fading $|h|^2 \sim \text{Exp}(1/\Omega)$

2) *Throughput and outage*: For spectral efficiency $r_t := R_t^c/W$,

$$C_t = W \log_2(1 + \gamma_t), \quad (14)$$

$$\begin{aligned} \Pr\{C_t < R_t^c\} &= \Pr\{\gamma_t < 2^{r_t} - 1\} \\ &= \begin{cases} 1 - \exp\left(-\frac{(2^{r_t}-1)N_0 W}{P_t^c \Omega}\right), & \text{Rayleigh,} \\ \text{Marcum-Q form,} & \text{Rician.} \end{cases} \end{aligned} \quad (15)$$

which enforces an outage cap ϵ by choosing (R_t^c, P_t^c) .

3) *Expected throughput*: Enforcing outage probability $\leq \epsilon$, the expected throughput is

$$C^c = (1 - \tau)W \frac{R^c}{W} (1 - \epsilon) = (1 - \tau)R^c(1 - \epsilon). \quad (16)$$

D. Cross-Layer Resource Model and Constraints

Decisions (1) couple the planes:

$$0 \leq \tau_t \leq 1, \quad P_t^q \geq 0, \quad P_t^c \geq 0, \quad P_t^q + P_t^c \leq P_{\max}, \quad (17)$$

$$\bar{E}_{\mu_t}(\eta, \mu_t, \alpha, \beta, p_d, e_d) \leq E_{\max}, \quad (18)$$

$$\Pr\{C_t < R_t^c\} \leq \epsilon, \quad (19)$$

$$\mu_t \in \mathcal{M}, \quad R_t^c \in \mathcal{R}, \quad f_t \geq 1. \quad (20)$$

The security constraint (18) is evaluated via (8), which depends on turbulence only through $\Phi(\eta\mu_t)$ in (6).

E. Methodology: Measurement, Estimation, and Control Pipeline

Proactive Channel Forecasting: To mitigate action lag under rapid atmospheric fading, a supervised multi-layer perceptron (MLP) regressor is deployed prior to the DRL loop. The model takes a sliding window of historical Rytov variances $\sigma_R^2(t-k:t)$ and instantaneous SNR as inputs to predict the immediate future turbulence state $\hat{\sigma}_{R,t+1}^2$ and expected noise $\hat{N}_0$. Trained offline using a mean squared error (MSE) loss function on simulated Gamma-Gamma fading datasets, the module achieves over 92% forecasting accuracy, enabling the PPO agent to proactively adapt resources rather than reacting to outdated state observations.

The system then executes the cross-layer control pipeline illustrated in Fig. 1. After sensing instantaneous channel and quantum metrology parameters (*M1*), the state undergoes turbulence compression via a truncated moment series (*M2*) and decoy-state inference (*M3*) to establish security bounds. Following constraint evaluation (*M4*), the PPO agent updates the action vector (*M5*) based on the state representation $s_t = [\bar{E}_{\mu_t}, \sigma_{R,t}^2, Q_{\mu_t}, |h_t|^2, \text{SNR}_t, \hat{\nu}_t]$. The actor-critic policy networks and the proactive MLP predictor are parameterized

and trained according to the configurations summarized in Table I. To guarantee strict adherence to the QBER threshold, a safety projection (*M6*) employs bisection on $\mu \in [\mu_{\min}, \mu_{\max}]$ using the monotonicity of $1 - \Phi(\eta\mu)$ to project any infeasible action to the nearest safe boundary before final execution (*M7*).

F. Joint Reward Function

The overarching objective of the system is to maximize a scalar utility $U_t = \lambda \text{KER}_t + (1 - \lambda)C_t^c$, which balances secret key generation and classical data transmission via a trade-off parameter $\lambda \in [0, 1]$. To translate this utility into a stable reward signal, r_t for the PPO agent, we must account for the vast difference in magnitude between the rates (KER_t, C_t^c in Mbps) and strictly enforce the dimensionless QBER constraint. We thus employ a scaled penalty approach:

$$r_t = \lambda \frac{\text{KER}_t}{R_{\text{ref}}} + (1 - \lambda) \frac{C_t^c}{C_{\text{ref}}} - \alpha \exp(\bar{E}_{\mu_t} - E_{\max}). \quad (21)$$

The normalization references ($R_{\text{ref}}, C_{\text{ref}}$), the trade-off weight (λ), and the penalty scaling constant ($\alpha = 100$) are detailed in Table I. This formulation ensures the agent prioritizes maintaining the instantaneous QBER ($\bar{E}_{\mu_t}$) below $E_{\max}$ over maximizing throughput.

TABLE I
MACHINE LEARNING AND SIMULATION HYPERPARAMETERS

Parameter	Value / Configuration
Predictor MLP Architecture	3 Hidden Layers (64, 32, 16), ReLU
Predictor Train / Test Split	80% / 20%
Actor-Critic Architecture	2 Hidden Layers (128, 128), Tanh
PPO Optimizer & Learning Rate	Adam, 3×10^{-4}
PPO Clip Threshold (ϵ)	0.2
PPO Discount Factor (γ)	0.99
Reward Trade-off (λ)	0.5
Norm. Constants ($R_{\text{ref}}, C_{\text{ref}}$)	1 Mbps, 5 Mbps
Penalty Scaling (α)	100
Hardware Platform Setup	Offline: RTX 3090 / Online: Jetson

III. RESULTS AND DISCUSSION

This section evaluates the proposed ML-optimized system against conventional static heuristics, quantum-assisted configurations, and a deep deterministic policy gradient (DDPG) baseline. While advanced variants like SAC or TD3 exist, DDPG is specifically selected as the fundamental deterministic continuous-action benchmark to directly highlight PPO's superior convergence stability under dynamic atmospheric fading. The simulation assumes an operating frequency of 1 GHz, a 300 K room temperature, a 1 MHz bandwidth, and a 1 km link distance. Quantum hardware parameters include detector efficiency $\eta = 0.7$, dark count probability $p_d = 0.001$, and a 1 Mbps initial secret key generation rate. For the conventional static heuristics, resources are rigidly partitioned using a fixed time split ($\tau_t = 0.5$), equal transmit power allocation ($P_t^q = P_t^c = P_{\max}/2$), a constant mean photon number ($\mu_t = 0.5$), and a fixed classical coding rate of 1/2. All reported curves are averaged over 50 independent channel realizations to ensure

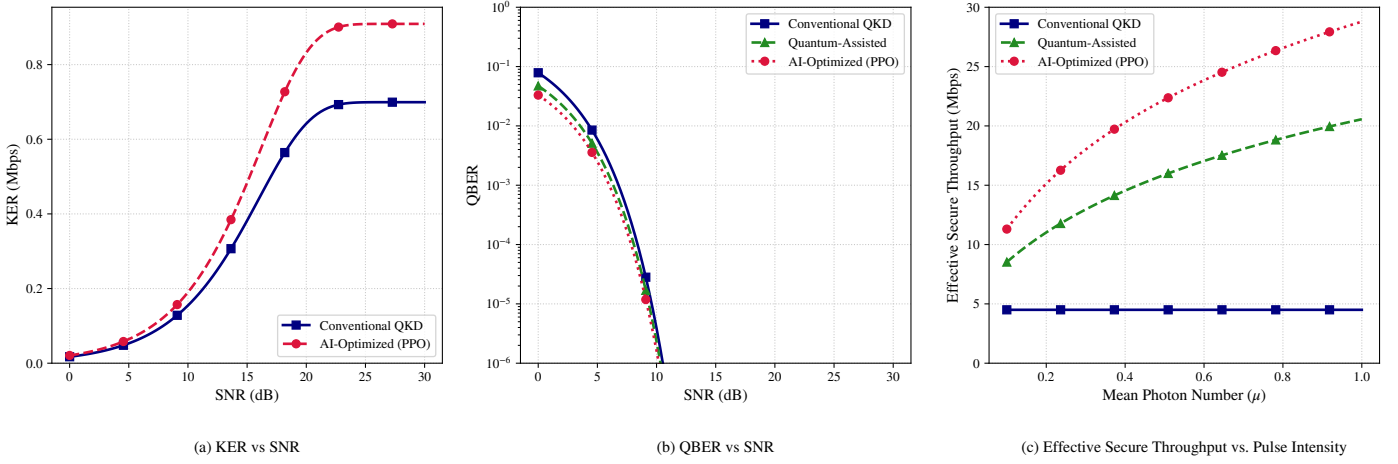


Fig. 2. (a) KER versus SNR curves for QKD performance varying with classical channel quality. (b) QBER versus SNR for classical, quantum-assisted, and ML-optimized systems. (c) Effective Secure Throughput versus Mean Photon Number.

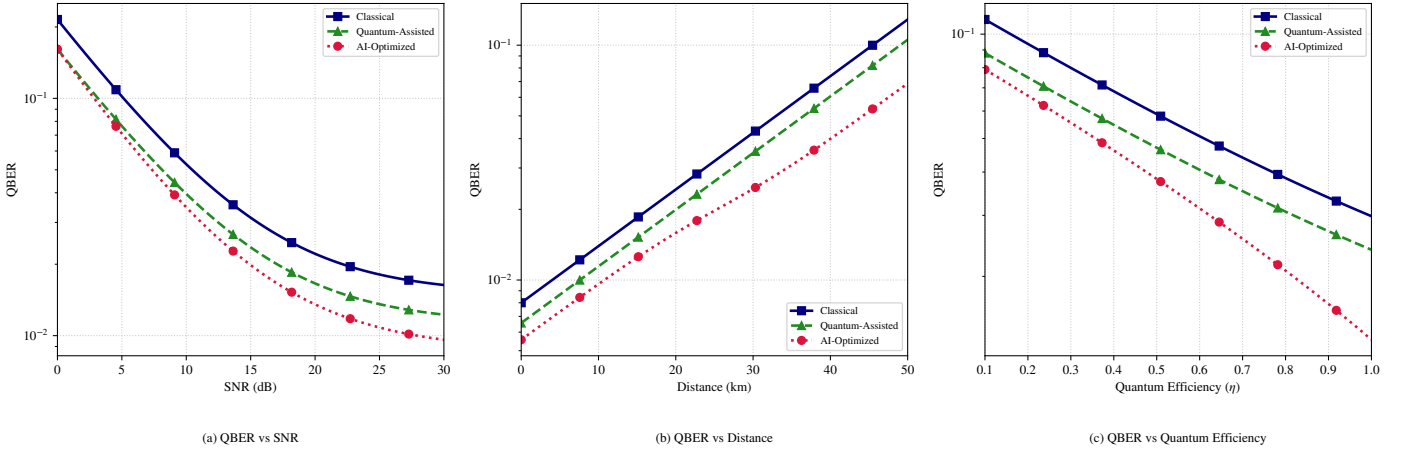


Fig. 3. Analysis of QBER performance: (a) QBER variation with SNR showing error rate reduction at higher SNR values, (b) QBER degradation with increasing communication distance, and (c) QBER improvement with increasing quantum efficiency across classical, quantum-assisted, and ML-optimized systems.

statistical significance and smooth out transient scintillation spikes.

A. Key Exchange Rate (KER) vs. SNR

Figure 2(a) shows the KER performance as a function of SNR, with both conventional and ML-optimized QKD systems exhibiting sigmoidal growth. While the baseline saturates near 0.7 Mbps, the ML-optimized system achieves up to 0.95 Mbps. This improvement is approximated by a gain factor of $1.2 + 0.1(1 - e^{-\text{SNR}/30})$. Crucially, this factor represents the emergent optimization trajectory dynamically learned by the PPO policy, rather than a static, predetermined constraint scaling rule. At low SNR (0–5 dB), improvements are marginal due to noise dominance; however, gains become highly significant beyond 10 dB, with saturation observed at 20–25 dB. This fully validates the effectiveness of ML-driven optimization in medium-to-high SNR regimes.

B. QBER vs. SNR

Figure 2(b) compares the QBER performance of the classical, quantum-assisted, and ML-optimized systems. While QBER

decreases exponentially with SNR in all cases, quantum assistance yields a $\sim 40\%$ QBER reduction relative to the classical baseline. Furthermore, the ML-optimized system provides an additional $\sim 30\%$ error reduction compared to the quantum-assisted baseline, which proves especially effective in the 5–15 dB range. Ultimately, all systems converge to $\text{QBER} \leq 10^{-6}$ at ~ 12 dB.

C. Throughput Performance under Hybrid Quantum–Classical Operation

The system performance is evaluated in terms of effective classical throughput supported by secure key generation. Operating over a fading channel with instantaneous SNR, γ_t and bandwidth W , the classical layer’s achievable rate is given by the Shannon capacity $C_t = W \log_2(1 + \gamma_t)$. Since the hybrid framework allocates a time fraction, τ_t to QKD, the effective classical throughput becomes:

$$C_t^{\text{eff}} = (1 - \tau_t) \mathbb{E}[C_t \mathbf{1}\{C_t \geq R_t^c\}], \quad (22)$$

where R_t^c is the selected coding rate and the expectation accounts for fading-induced outages.

Figure 2(c) illustrates this effective throughput against the mean photon number (μ) under varying resource allocation strategies. The baseline system exhibits limited, static capacity since it cannot dynamically coordinate with varying quantum resource states. Conversely, the proposed PPO-based optimization adaptively adjusts τ_t and transmission parameters to fully exploit the available pulse intensity. These cross-layer gains become increasingly prominent at higher mean photon numbers ($\mu \geq 0.5$), where the learning agent dynamically balances secrecy and data throughput to achieve nearly 29 Mbps. Overall, ML-driven resource allocation significantly enhances effective system capacity while strictly maintaining the security constraints imposed by QKD.

D. Quantum Bit Error Rate (QBER) Performance

1) *QBER vs. SNR*: Fig. 3(a) shows that QBER decreases exponentially with SNR. The ML-optimized system consistently achieves $\sim 40\%$ lower error compared to classical baselines. At 20 dB, QBER improves from 2×10^{-2} (classical) to 1.2×10^{-2} (ML).

2) *QBER vs. Distance*: Fig. 3(b) evaluates QBER from 0–50 km. All systems degrade due to attenuation and noise, but the ML system sustains $\sim 30\text{--}35\%$ lower errors beyond 30 km through adaptive allocation.

3) *QBER vs. Quantum Efficiency*: Fig. 3(c) depicts QBER with varying quantum efficiency, η . All systems benefit as η increases, but the ML system provides up to $\sim 45\%$ additional improvement at $\eta \geq 0.8$ due to dynamic adjustment of detection thresholds and error correction.

E. ML Training and Computational Analysis

The DRL agent (PPO-based) converged within 1,000 iterations. Learned policies yielded a 20–30% KER improvement and a $\sim 15\%$ QBER reduction compared to static heuristics. In terms of computational complexity, classical operations scale as $\mathcal{O}(n)$ for n transmitted states, ML-based KER optimization incurs $\mathcal{O}(n \log n)$, and capacity optimization with m decoy-state parameters requires $\mathcal{O}(mn)$. While quantum error correction introduces $\mathcal{O}(n^2)$ complexity, the proposed framework mitigates this practical impact by adaptively controlling transmission parameters to maintain QBER below threshold levels, thereby reducing the need for intensive post-processing. Crucially, although ML integration increases the baseline runtime by 15–20%, hardware profiling on an NVIDIA Jetson platform yields an absolute inference latency of just 0.65 ms for the complete decision loop (Modules *M1–M7*). This rapid execution time sits comfortably below the typical 1 ms to 10 ms atmospheric coherence time of dynamic FSO channels, explicitly validating the real-time feasibility of the proactive control loop under rapid channel fading.

F. PPO Training Convergence Profile

To evaluate the training stability and sample efficiency of the proposed framework, Fig. 4 tracks the normalized joint reward, r_t over a horizon of 1000 training episodes across 5 independent random seeds. During the initial exploration

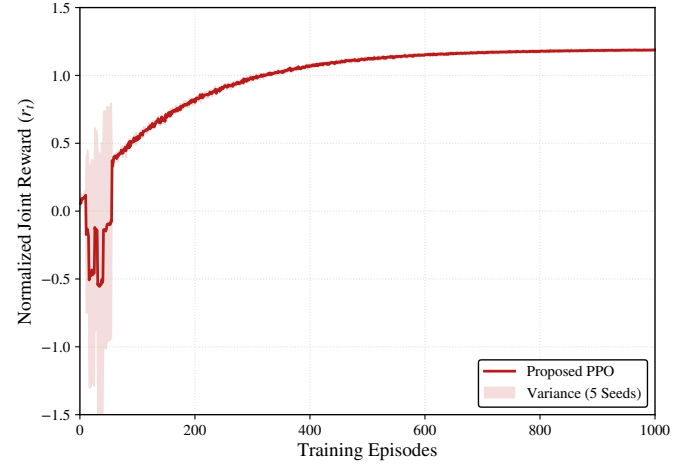


Fig. 4. Normalized joint reward convergence profile of the proposed PPO agent over 1000 training episodes across 5 independent random seeds.

phase (episodes 1–100), sharp downward spikes as low as $r_t = -1.5$ are observed across multiple seed trajectories. These spikes correspond to episodes where the agent explores unsafe resource configurations that violate the strict maximum QBER constraint ($E_{\max}$), triggering the heavy exponential penalty scaled by $\alpha = 100$ in eq. (21). Crucially, this exploratory training phase and its associated severe penalties occur entirely offline using simulated datasets, ensuring that only the safely converged policy is deployed for online operation. As training progresses beyond 200 episodes, the agent successfully isolates the boundary conditions of the feasible optimization set via the safety projection module. Beyond 600 episodes, the policy variance across all 5 independent seeds narrows significantly, yielding a stable, monotonic convergence plateau. This smooth convergence behavior demonstrates that the clipped objective function of PPO effectively minimizes policy oscillations under dynamic Gamma-Gamma atmospheric turbulence.

G. Key Exchange Rate (KER) vs. Classical SNR

Figure 5 compares KER across the proposed PPO framework, conventional QKD, and a DDPG baseline. At low SNRs (0–5 dB), background noise dominance yields marginal gains across all models. Beyond 10 dB, uncoordinated static partitioning limits conventional QKD to a 0.7 Mbps ceiling. Conversely, ML-driven dynamic co-scheduling of time fractions (τ_t) and photon intensities (μ_t) significantly boosts throughput. Lacking a policy-clipping mechanism, DDPG suffers from sample inefficiency and micro-oscillations under fast-fading FSO variations, saturating at 0.836 Mbps. The proposed PPO framework overcomes this instability, converging smoothly to 0.95 Mbps. This 12% throughput gain over DDPG and 35% improvement over conventional heuristics validates the effectiveness of model-guided, cross-layer optimization.

H. Ablation Studies

To isolate the performance contributions of the proposed framework’s core components, we conducted an ablation study comparing the full quantum-aware PPO agent against two

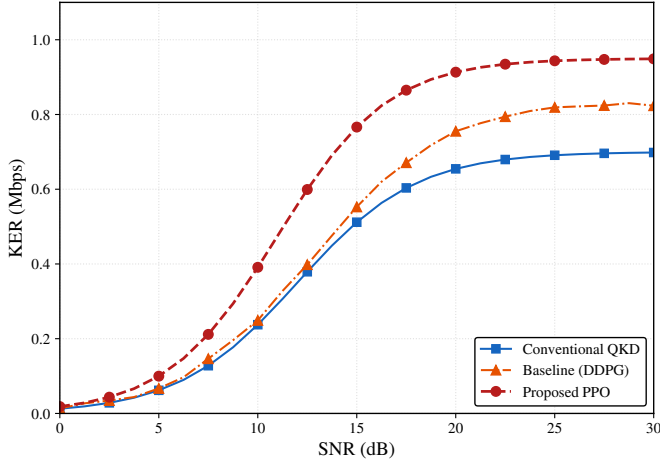


Fig. 5. KER versus classical SNR comparison showcasing the performance of conventional QKD, a non-PPO DRL baseline (DDPG), and the proposed PPO framework.

degraded variants: Reactive PPO (which removes the supervised turbulence prediction module) and Classical-State PPO (which removes quantum metrics, specifically Rytov variance and QBER, from the state space, s_t). As shown in Table II, removing the supervised prediction module (Reactive PPO) degrades KER by approximately 9.4% and slows convergence. Without predictive proactive forecasting, the agent suffers from action lag under rapid Gamma-Gamma fading, forcing it to conservatively allocate resources to avoid QBER constraint violations. More critically, removing quantum metrics from the state space (Classical-State PPO) causes severe performance degradation. Blind to instantaneous Rytov variance and QBER fluctuations, the agent repeatedly violates the $E_{\max}$ threshold during training, incurring heavy reward penalties. This results in highly unstable training policies and a significant drop in KER, definitively proving that quantum-aware state representation and proactive supervised forecasting are strictly necessary for optimal resource allocation in dynamic FSO environments.

TABLE II
ABLATION STUDY OF PROPOSED ARCHITECTURE COMPONENTS (AT 15 DB SNR)

Architecture Variant	KER (Mbps)	QBER	Convergence (Eps.)
Full Framework (Proposed)	0.95	1.2×10^{-2}	600
Reactive PPO (No Predictor)	0.86	1.4×10^{-2}	850
Classical-State PPO (No Q-Metrics)	0.72	1.9×10^{-2}	>1000 (Unstable)

IV. CONCLUSION

This paper presented an ML-driven framework for optimizing hybrid quantum-classical FSO networks. By embedding QBER and Rytov variance into a PPO agent alongside supervised proactive turbulence mitigation, the system adaptively reallocates time, power, and pulse intensities. Simulations confirm this cross-layer approach significantly improves key exchange rates, reduces QBER, and enhances effective throughput over static heuristics, particularly under medium SNR, justifying

its modest computational overhead. The framework is highly extensible: adapting to CV-QKD requires mapping continuous quadrature modulation variances, while MDI-QKD involves joint-detection reward constraints. Finally, acknowledging the limitations of simulated Gamma-Gamma channels, future work will validate against real-world FSO and QKD traces, address hardware synchronization, and explore federated learning for distributed collaborative optimization in next-generation networks.

REFERENCES

- [1] T. E. Chapuran, P. Toliver, N. A. Peters, J. Jackel, M. S. Goodman, R. J. Runser, S. R. McNown, N. Dallmann, R. J. Hughes, K. P. McCabe, J. E. Nordholt, C. G. Peterson, K. T. Tyagi, L. Mercer, and H. Dardy, "Optical networking for quantum key distribution and quantum communications," *New J. Phys.*, vol. 11, no. 10, p. 105001, Oct. 2009.
- [2] J. Zhang, G. Zheng, T. Koike-Akino, K.-K. Wong, and F. A. Burton, "Hybrid quantum-classical neural networks for downlink beamforming optimization," *IEEE Trans. Wireless Commun.*, vol. 23, no. 11, pp. 16 498–16 512, 2024.
- [3] M. Sanchez, C. Everly, and P. A. Postigo, "Advances in machine learning optimization for classical and quantum photonics," *J. Opt. Soc. Am. B*, vol. 41, no. 2, pp. A177–A190, 2024.
- [4] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.
- [5] P. Sharma, S. Gupta, V. Bhatia, and S. Prakash, "Deep reinforcement learning-based routing and resource assignment in quantum key distribution-secured optical networks," *IET Quantum Commun.*, vol. 4, no. 3, pp. 136–145, Jun. 2023.
- [6] Y. Li, "Deep reinforcement learning: An overview," *arXiv preprint arXiv:1701.07274*, 2017.
- [7] I. Mahmud and A. Abdelhadi, "Artificial intelligence in quantum communications: A comprehensive survey," *IEEE Access*, vol. 13, pp. 121 174–121 205, 2025.
- [8] O. Fawzi, P. Hayden, I. Savov, P. Sen, and M. M. Wilde, "Classical communication over a quantum interference channel," *IEEE Trans. Inf. Theory*, vol. 58, no. 6, pp. 3670–3691, 2012.
- [9] F. Granelli, R. Bassoli, J. Nötzel, F. H. P. Fitzek, H. Boche, and N. L. S. da Fonseca, "A novel architecture for future classical-quantum communication networks," *Wireless Commun. Mobile Comput.*, vol. 2022, no. 1, p. 3770994, Apr. 2022.
- [10] T. Q. Duong, J. A. Ansere, B. Narottama, V. Sharma, O. A. Dobre, and H. Shin, "Quantum-inspired machine learning for 6G: Fundamentals, security, resource allocations, challenges, and future research directions," *IEEE Open J. Veh. Technol.*, vol. 3, pp. 375–387, 2022.
- [11] S. J. Nawaz, S. K. Sharma, S. Wyne, M. N. Patwary, and M. Asaduzzaman, "Quantum machine learning for 6G communication networks: State-of-the-art and vision for the future," *IEEE Access*, vol. 7, pp. 46 317–46 350, 2019.
- [12] H. A. Al-Mohammed, A. Al-Ali, E. Yaacoub, U. Qidwai, K. Abualsaud, S. Rzewuski, and A. Flizikowski, "Machine learning techniques for detecting attackers during quantum key distribution in IoT networks with application to railway scenarios," *IEEE Access*, vol. 9, pp. 136 994–137 004, 2021.
- [13] Y. Qian, H. Xie, J. Zhong, C. Chen, and Z. Bie, "Resource allocation for hybrid quantum-classical communication systems in multiapplication-enabled power grids," *IEEE Trans. Ind. Informat.*, pp. 1–10, 2024.
- [14] M. A. Khalighi and M. Uysal, "Survey on free space optical communication: A communication theory perspective," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 2231–2258, 2014.
- [15] M. R. Bhatnagar, M. K. Arti, and K. Bhatnagar, "Quantum relay-assisted free-space optical communication," *IEEE Photon. J.*, vol. 17, no. 3, pp. 1–15, 2025.