DeepXplain: XAI-Guided Autonomous Defense Against Multi-Stage APT Campaigns

Trung V. Phan and Thomas Bauschert
Chair of Communication Networks, Technische Universität Chemnitz, 09126 Chemnitz, Germany
Email: trung.phan-van@etit.tu-chemnitz.de, thomas.bauschert@etit.tu-chemnitz.de
This paper is currently under review for IEEE GLOBECOM 2026.

***Abstract*—Advanced Persistent Threats (APTs) are stealthy, multi-stage attacks that require adaptive and timely defense. While deep reinforcement learning (DRL) enables autonomous cyber defense, its decisions are often opaque and difficult to trust in operational environments. This paper presents *DeepXplain*, an explainable DRL framework for stage-aware APT defense. Building on our prior DeepStage model [1], DeepXplain integrates provenance-based graph learning, temporal stage estimation, and a unified XAI pipeline that provides structural, temporal, and policy-level explanations. Unlike post-hoc methods, explanation signals are incorporated directly into policy optimization through evidence alignment and confidence-aware reward shaping. To the best of our knowledge, DeepXplain is the first framework to integrate explanation signals into reinforcement learning for APT defense. Experiments in a realistic enterprise testbed show improvements in stage-weighted F1-score (0.887 to 0.915) and success rate (84.7% to 89.6%), along with higher explanation confidence (0.86), improved fidelity (0.79), and more compact explanations (0.31). These results demonstrate enhanced effectiveness and trustworthiness of autonomous cyber defense.**

***Index Terms*—Advanced Persistent Threat (APT), Deep Reinforcement Learning (DRL), Autonomous Cyber Defense, Provenance Graph Embedding, Explainable AI (XAI).**

I. Introduction

Advanced Persistent Threats (APTs) [2] remain one of the most challenging classes of cyber attacks against enterprise networks. Unlike opportunistic malware, APT campaigns are stealthy, low-and-slow, and multi-stage, evolving through reconnaissance, initial compromise, privilege escalation, lateral movement, command-and-control, and exfiltration. Their distributed and temporally extended nature makes them difficult to detect and contain using traditional rule-based or signature-based defenses, which lack the ability to reason over causal system interactions and long-term attack progression.

Recent advances in deep reinforcement learning (DRL) have enabled the development of autonomous cyber-defense agents capable of adapting to dynamic threats. In our prior work, we proposed *DeepStage* [1], a stage-aware autonomous defense framework that models enterprise activity using fused provenance graphs and learns mitigation policies via DRL. By combining graph-based representation learning and temporal stage estimation, DeepStage improves defense effectiveness against multi-stage APT campaigns. However, similar to many DRL-based systems, it operates as a black box: while it produces effective decisions, it does not provide interpretable explanations for its predictions or actions.

This limitation is particularly critical in security operations. Defense actions such as host isolation, credential revocation, and traffic blocking can incur significant operational disruption and therefore require validation by human analysts. Without interpretable evidence, autonomous defense policies remain difficult to trust, audit, and deploy in practice. Recent advances in explainable reinforcement learning (XRL) [3] emphasize the need for transparency in sequential decision-making systems, particularly in safety-critical domains. These studies highlight that DRL agents often lack interpretability, which limits their adoption in real-world settings where reliability and accountability are essential. In parallel, explainable graph neural network (GNN) methods [4] have emerged as effective tools for identifying influential nodes, edges, and subgraphs that drive model predictions. Such capabilities provide a strong foundation for interpreting provenance-based security analytics. However, existing approaches primarily focus on static prediction tasks and do not establish a direct connection between graph-level explanations and sequential decision-making in reinforcement learning [3], [4].

In this paper, we propose *DeepXplain*, an XAI-guided extension of DeepStage [1] for autonomous defense against multi-stage APT campaigns. DeepXplain introduces a unified explanation pipeline that captures structural evidence from provenance graphs, temporal attribution over attack evolution, and feature-level policy attribution for defense decisions. Unlike conventional post-hoc approaches, these explanation signals are incorporated directly into reinforcement learning through evidence-alignment regularization and confidence-aware reward shaping. This tightly couples explanation with policy learning, enabling the agent to make decisions that are both effective and interpretable. We evaluate DeepXplain in a realistic enterprise testbed using CALDERA-driven APT scenarios [5]. Experimental results show that DeepXplain improves the average stage-weighted F1-score from 0.887 to 0.915 and increases mitigation success rate from 84.7% to 89.6%. In addition, it produces more reliable and concise explanations, achieving higher explanation confidence, compactness and fidelity compared to the post-hoc baseline.

II. Related Work

A. Explainable Reinforcement Learning

Recent research on explainable reinforcement learning (XRL) has emphasized the need to make sequential decision-

making policies interpretable, particularly in safety- and mission-critical domains. Milani *et al.* [3] provide a recent comprehensive survey of XRL and organize the literature around feature attribution, policy summarization, and explanation generation. Their study highlights that most XRL methods remain domain-agnostic and are rarely integrated into the optimization objective itself. Despite this progress, the application of XRL to cyber defense remains limited. Existing autonomous cyber-defense agents typically emphasize mitigation performance, policy convergence, or robustness, but do not explicitly constrain learned policies to align with interpretable evidence. As a result, current RL-based cyber-defense systems often remain difficult to validate and trust in operational settings.

B. Explainable Graph Neural Networks

With the widespread adoption of graph neural networks (GNNs) for structured data analysis, explainability has become a critical research direction for understanding model decisions. Existing methods aim to identify the most influential nodes, edges, or subgraphs that contribute to a given prediction. For instance, GNNExplainer [4] learns soft masks over graph structures to maximize the mutual information between a subgraph and the model output, providing instance-level explanations. Recent surveys further systematize this area. Li *et al.* [6] categorize GNN explanation methods and highlight key evaluation criteria, including correctness, robustness, usability, and compactness. Similarly, Dai *et al.* [7] emphasize that explainability is a fundamental component of trustworthy GNNs, alongside robustness and privacy. Despite these advances, existing approaches are primarily designed for static tasks such as node classification or graph prediction. They do not directly address dynamic and security-critical settings, where graph structures evolve over time and decisions must be taken sequentially. This limitation motivates the need for a unified framework that bridges provenance-graph explanation with policy learning.

C. Autonomous Cyber Defense and RL-Based Mitigation

Autonomous cyber defense has recently attracted significant attention as defenders seek to respond to machine-speed attacks using learning-based agents. Le *et al.* [8] propose a recent RL-based framework for automated APT defense using attack-graph risk-based situation awareness. Their approach shows that reinforcement learning can improve adaptive mitigation under staged attack conditions, but it relies on abstract attack-graph representations and does not expose interpretable evidence for individual decisions. Similarly, recent work on entity-based reinforcement learning for autonomous cyber defense [9] explores more realistic defense states, but still focuses primarily on action effectiveness rather than explainability. These findings motivate the need for defense agents whose decisions are not only effective but also transparent and evidence-driven.

D. Provenance-Based Security Analytics

Recent provenance-based intrusion detection systems have shown strong potential for detecting stealthy and multi-stage attacks. FLASH [10] is a recent representative system that applies graph representation learning to system provenance graphs for scalable intrusion detection. Its results demonstrate the effectiveness of provenance-aware GNN models in capturing causal attack structure beyond flat log features. However, the focus remains on detection rather than autonomous mitigation. More recently, Bilot *et al.* [11] present a comprehensive analysis of provenance-based intrusion detection systems and show that, despite strong reported detection performance, many existing systems face practical challenges related to reproducibility, deployment realism, and analyst usability. This is particularly important for our setting: provenance-based systems can provide rich causal evidence, but without interpretable and action-oriented reasoning they remain difficult to operationalize in autonomous defense workflows.

E. Research Gap

Taken together, recent work reveals three key gaps. First, XRL methods improve policy transparency but are rarely integrated with cyber-defense agents operating in partially observable, graph-structured environments. Second, explainable GNN research provides strong tools for extracting structural evidence, but does not generally connect that evidence to downstream defense policies. Third, autonomous cyber-defense and provenance-based IDS systems have advanced detection and mitigation capabilities, yet recent analyses show persistent challenges in trust, deployment realism, and analyst interpretability. These limitations motivate the need for a unified framework that links structural explanations with sequential decision making for interpretable APT defense.

III. XAI-Guided Autonomous APT Defense

This section presents *DeepXplain*, an XAI-guided extension of the DeepStage framework [1] for autonomous defense against multi-stage APT campaigns, as illustrated in Figure 1. While DeepStage enables adaptive defense through deep reinforcement learning, its decision process remains largely opaque, limiting trust and operational deployment.

DeepXplain addresses this limitation by tightly integrating explainable artificial intelligence (XAI) into both inference and decision making. Specifically, it unifies provenance-graph explanation, temporal attribution, and policy-level interpretation, and incorporates these explanation signals directly into policy optimization. This design enables the learned defense policy to be not only effective but also evidence-driven, interpretable, and reliable in security-critical environments.

A. System Model

Enterprise telemetry from host and network monitoring systems is transformed into a provenance graph $G_t = (V_t, E_t)$, where nodes V_t represent system entities (e.g., processes, files, sockets), and edges E_t encode causal relationships such as

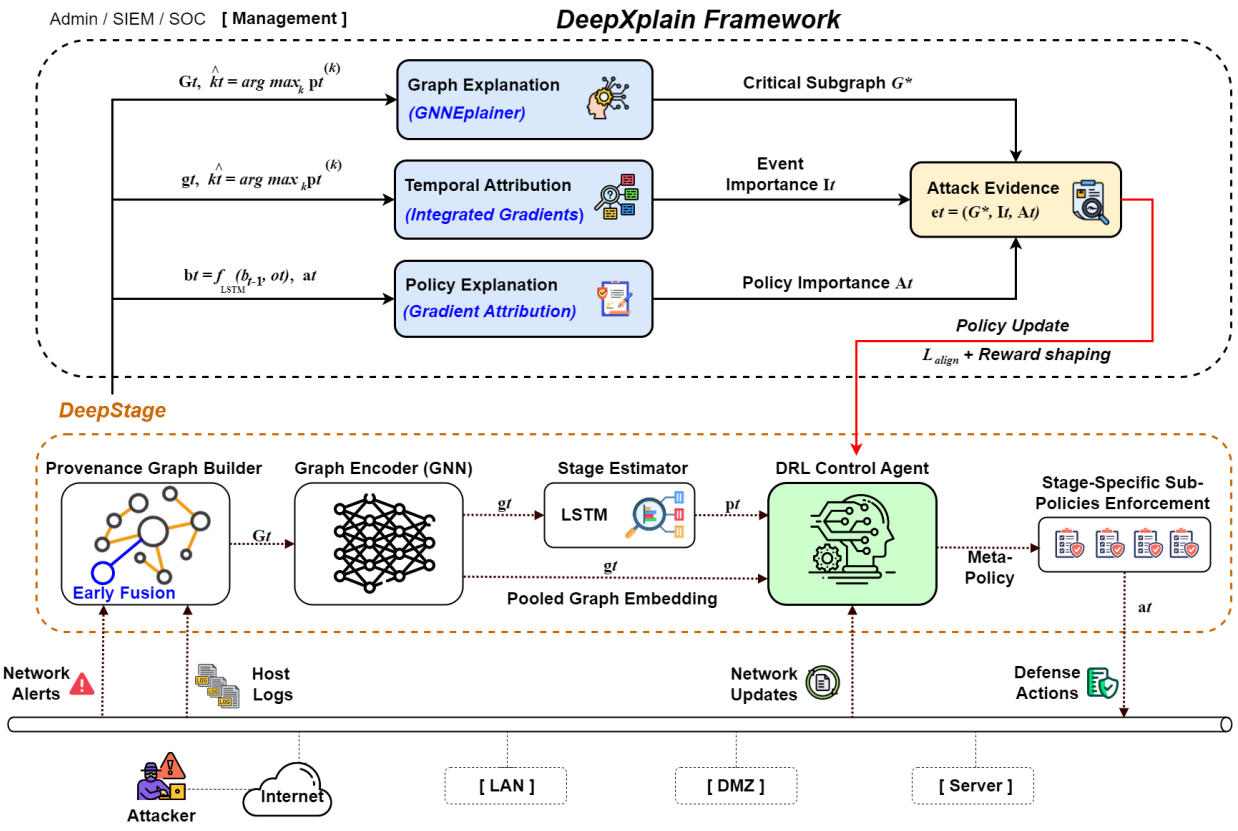

Fig. 1. Data and control flow of the DeepXplain framework. The red feedback path highlights how explanation signals are incorporated into policy optimization.

process spawning or network communication. This representation captures both structural dependencies and event causality across the system.

A graph neural network (GNN) encodes the provenance graph as $g_t = f_{\text{GNN}}(G_t)$, where $g_t \in \mathbb{R}^d$. The embedding g_t summarizes complex interactions within the system into a compact representation suitable for learning.

To capture temporal evolution, the sequence of embeddings is processed by a Stage Estimator: $p_t = f_{\text{LSTM}}(g_1, \ldots, g_t)$, where $p_t \in \mathbb{R}^K$. The vector p_t represents the probability distribution over K APT stages, enabling temporal reasoning about attack progression.

Since the true system state is partially observable, we adopt a POMDP formulation. The belief state is updated as $b_t = f_{\text{LSTM}}(b_{t-1}, o_t)$ with observation $o_t = [g_t, p_t, a_{t-1}]$, where b_t integrates historical observations and actions to approximate the latent system state. The policy then selects a defense action $a_t \sim \pi_\theta(a|b_t)$, allowing context-aware mitigation strategies.

B. XAI Pipeline

To enhance interpretability, DeepXplain introduces an XAI module that operates on intermediate representations of the model. At each time step, the module receives

$$\mathcal{X}_t = \{G_t, g_t, p_t, b_t, a_t\}, \tag{1}$$

which jointly capture structural, temporal, and decision-level information.

The module produces an explanation signal

$$e_t = (G_t^*, I_t, A_t), \tag{2}$$

where G_t^* identifies critical graph structures, I_t highlights important time steps, and A_t explains the selected action. This unified representation enables comprehensive reasoning across multiple dimensions of the defense process.

C. Graph-Based Explanation

To explain attack-stage predictions, we identify a subgraph that preserves the model prediction:

$$G_t^* = \arg\max_{G' \subseteq G_t} P(\hat{k}_t | G'), \quad \hat{k}_t = \arg\max_k p_t^{(k)}. \tag{3}$$

This formulation seeks the minimal structural evidence sufficient to maintain the predicted stage.

We adopt GNNExplainer [4], which learns soft masks

$$M_V \in [0,1]^{|V_t|}, \quad M_E \in [0,1]^{|E_t|}, \tag{4}$$

assigning importance scores to nodes and edges. These scores highlight critical attack patterns such as suspicious process chains or lateral movement paths.

D. Temporal Attribution

APT attacks evolve over time, making it essential to identify critical temporal events. We compute temporal attribution as

$$I_i = \left| \frac{\partial P(\hat{k}_t)}{\partial g_i} \right|_1. \tag{5}$$

This measures how strongly each past embedding influences the current prediction. The normalized importance $\tilde{I}_i = \frac{I_i}{\sum_j I_j}$ forms a probability distribution over time steps, enabling identification of key transitions in the attack lifecycle.

E. Explainable Defense Policy

To interpret defense decisions, we analyze the sensitivity of the policy to the belief state:

$$A_i = \left| \frac{\partial \pi(a_t|b_t)}{\partial b_t^{(i)}} \right|. \tag{6}$$

This reveals which features—such as high attack-stage probability or anomalous activity—drive the selected action. The normalized attribution is defined as $\phi_{\text{policy}} = \frac{A_t}{\|A_t\|_1}$, which provides a comparable importance distribution.

To align structural explanations with policy reasoning, we project graph explanations into belief space:

$$g_t^* = \sum_{v \in V_t} w(v) h_v, \tag{7}$$

where $w(v)$ is node importance and h_v is the node embedding. We then compute $\phi_{\text{XAI}} = \frac{W_g g_t^* + W_p p_t}{\|W_g g_t^* + W_p p_t\|_1}$, which integrates structural evidence and stage information into the same feature space as the policy, enabling direct comparison.

F. XAI-Guided Policy Optimization

We incorporate explanation signals into policy learning through an alignment loss:

$$\mathcal{L}_{align} = \|\phi_{\text{policy}} - \phi_{\text{XAI}}\|_2^2, \tag{8}$$

which encourages the policy to focus on features supported by interpretable evidence.

To quantify explanation reliability, we define

$$\text{Conf}(e_t) = \alpha C_{\text{graph}} + \beta C_{\text{temp}} + \gamma C_{\text{policy}}, \tag{9}$$

where each term captures a different aspect of explanation quality. Here, $C_{\text{graph}} = \frac{\sum_{v \in V_t^*} w(v)}{\sum_{v \in V_t} w(v)}$ measures structural completeness, $C_{\text{temp}} = 1 - \frac{-\sum_i \tilde{I}_i \log \tilde{I}_i}{\log t}$ measures temporal concentration, and $C_{\text{policy}} = \frac{\phi_{\text{policy}}^\top \phi_{\text{XAI}}}{\|\phi_{\text{policy}}\|_2 \|\phi_{\text{XAI}}\|_2}$ measures consistency between explanation and policy.

Finally, the augmented objective is

$$J(\theta) = J_{RL}(\theta) - \lambda_1 \mathcal{L}_{align} + \lambda_2 \text{Conf}(e_t), \tag{10}$$

where λ_1 enforces evidence alignment and λ_2 rewards reliable explanations. In practice, this objective is optimized using PPO, where the augmented reward replaces the standard reward during policy updates. This formulation tightly integrates explanation and decision making, enabling DeepXplain to learn policies that are not only effective but also interpretable and robust against complex multi-stage APT behaviors.

IV. Performance Evaluation

This section evaluates the proposed *DeepXplain* framework with respect to both *defense effectiveness* and *explanation quality*. Since DeepXplain is built on top of DeepStage, we inherit the same enterprise testbed, telemetry pipeline, attack scenarios, and defense action space used in our prior work [1]. Unless otherwise stated, all components unrelated to explainability follow the same configuration as DeepStage to isolate the impact of the proposed XAI-guided extensions.

A. Experimental Setup

1) Enterprise Testbed and Data Sources: Following DeepStage [1], experiments are conducted in a realistic enterprise testbed logically segmented into four zones: local area network (LAN), demilitarized zone (DMZ), server zone, and management zone. Host-level telemetry is collected from endpoint monitoring tools (e.g., Auditd), while network-level events are captured by Zeek-based sensors. These events are normalized and fused into provenance graphs, where nodes represent system entities and edges encode causal interactions.

To emulate realistic multi-stage APT campaigns, we use CALDERA-driven adversarial playbooks [5] covering reconnaissance, initial compromise, privilege escalation, lateral movement, command-and-control, and exfiltration. Benign background activities are executed concurrently to create realistic noise and increase detection difficulty. The same attack-stage definitions, telemetry collection interval, and graph construction window used in DeepStage are retained here to ensure a fair comparison. All reported results are averaged over 10 independent runs.

2) Model Configuration: DeepXplain reuses the core DeepStage backbone, including the GNN encoder, LSTM-based Stage Estimator, and PPO-based defense policy. The graph encoder produces an embedding $g_t \in \mathbb{R}^d$, and the Stage Estimator outputs $p_t \in \mathbb{R}^K$. The belief state $b_t \in \mathbb{R}^{d_b}$ is updated recurrently using the same observation structure as DeepStage.

For the XAI components, the graph explanation module is implemented using GNNExplainer [4]. The explainer learns node and edge masks for each graph instance over 100 optimization steps with learning rate 10^{-2}. We retain the top-m explanation nodes after masking, where m is selected to preserve at least 90% of the cumulative node importance in G_t^*. Temporal attribution is computed using gradient-based sensitivity over the sequence of graph embeddings, and policy attribution is computed from the policy network gradients with respect to the belief-state features.

3) XAI-Guided Optimization Parameters: The XAI-guided objective introduced in Section III includes an alignment regularization term and a confidence-based reward term in Equation (10). In our experiments, the alignment coefficient is set to $\lambda_1 = 0.1$, which provides a moderate constraint encouraging policy attention to align with explanation-derived evidence while preserving policy flexibility. The explanation confidence weight is set to $\lambda_2 = 0.05$, allowing confidence-aware reward shaping without overwhelming the original defense objective. These values are chosen from the ranges

TABLE I
PER-STAGE F1-SCORE AND OVERALL DEFENSE EFFECTIVENESS COMPARISON.

Method	Recon	InitAcc	PrivEsc	LatMov	C2	Exfil	Avg-F1	Success Rate (%)
Risk-Aware DRL	0.75	0.73	0.70	0.71	0.76	0.72	0.728	68.5
DeepStage	0.91	0.88	0.85	0.87	0.92	0.89	0.887	84.7
DeepXplain	**0.93**	**0.91**	**0.89**	**0.90**	**0.94**	**0.92**	**0.915**	**89.6**

discussed in Section III, namely $\lambda_1 \in [0.01, 0.5]$ and $\lambda_2 \in [0.01, 0.3]$, and were found to provide stable training. For the explanation confidence score in Equation (9), we set $(\alpha, \beta, \gamma) = (0.4, 0.2, 0.4)$. This weighting prioritizes structural completeness and policy-evidence consistency, while still accounting for temporal concentration.

4) Baselines: We compare DeepXplain against the following baselines:

- *Risk-Aware DRL* [12]: A reinforcement learning-based defense framework that leverages attack-graph risk modeling to minimize cumulative security risk.
- *DeepStage* [1]: The original stage-aware autonomous defense framework without XAI guidance.
- *DeepStage + Post-hoc XAI*: DeepStage augmented with explanation modules for interpretability only, without incorporating explanation signals into policy learning.

B. Results Analysis

1) Defense Effectiveness: Table I summarizes both the per-stage and overall defense effectiveness of all methods under the same evaluation setting as DeepStage. DeepXplain consistently achieves the best performance across all APT stages, attaining the highest average stage-weighted F1-score of 0.915, compared to 0.887 for DeepStage and 0.728 for Risk-Aware DRL. The improvement is uniform across all stages and is particularly significant during privilege escalation and lateral movement, where accurate situational awareness and timely intervention are critical to preventing attack propagation. These results highlight the benefit of incorporating explanation-guided signals, which enable the agent to focus on causally relevant system behaviors rather than spurious correlations. Notably, these gains are obtained despite the already strong performance of DeepStage, indicating that explanation-aware learning contributes beyond representation learning alone.

The comparison with Risk-Aware DRL further demonstrates the limitations of static, risk-oriented state representations. While risk-aware approaches capture global attack surfaces, they rely on coarse attack-graph abstractions and lack the fine-grained, temporal, and causal context necessary to model multi-stage APT behavior. As a result, their performance degrades in later stages of the attack, where precise reasoning over system dynamics is essential.

Compared with DeepStage, DeepXplain achieves a notable improvement in overall mitigation effectiveness, increasing the success rate from 84.7% to 89.6%. The mitigation success rate is defined as the percentage of attack episodes in which the defense agent prevents the APT from reaching critical stages such as command-and-control or exfiltration. These

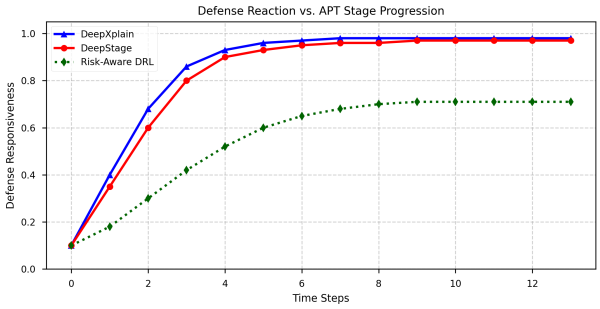

Fig. 2. Defense reaction versus APT stage progression.

gains indicate that integrating XAI into the learning process not only enhances interpretability but also leads to more reliable and efficient defense decisions. By aligning policy learning with interpretable evidence, DeepXplain enables the agent to prioritize compact, temporally consistent, and causally meaningful signals, thereby improving both robustness and response quality in complex APT scenarios.

2) Defense Responsiveness: Figure 2 illustrates the defense responsiveness across the APT lifecycle. DeepXplain consistently reacts faster and maintains higher responsiveness than all baselines. At early attack transitions ($t = 3$), DeepXplain achieves a responsiveness of 0.86, outperforming DeepStage (0.80, +7.5%) and Risk-Aware DRL (0.42, +104.8%). During critical escalation ($t = 5$), it reaches 0.96, compared to 0.93 for DeepStage (+3.2%) and 0.60 for Risk-Aware DRL (+60.0%). At convergence, DeepXplain stabilizes at approximately 0.98, exceeding DeepStage (0.97) and significantly outperforming Risk-Aware DRL (0.71, +38.0%). The advantage is most evident during the transition from privilege escalation to lateral movement, where rapid containment is crucial. This improvement is driven by explanation-guided policy optimization, which enables earlier detection of stage transitions and more timely, evidence-consistent responses.

3) Explanation Quality: Figure 3 evaluates the quality of the generated explanations from two complementary perspectives: confidence, compactness and fildelity. DeepXplain consistently outperforms DeepStage + Post-hoc XAI across all metrics. Specifically, DeepXplain achieves a higher explanation confidence score (0.86 vs. 0.71, +21.1%), indicating that the extracted evidence is more consistent and reliable. Compactness measures the conciseness of explanations, defined as the proportion of important graph elements required to explain a decision. Lower compactness values indicate that the expla-

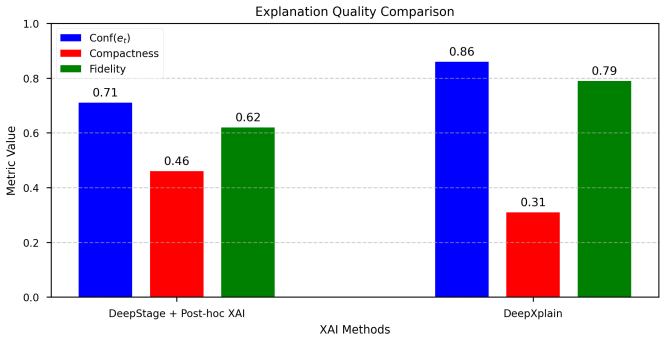

Fig. 3. Explanation quality comparison.

TABLE II
ABLATION STUDY ON XAI-GUIDED OPTIMIZATION.

Variant	Avg. Stage-F1	Conf(e_t)
DeepXplain w/o $\mathcal{L}_{align}$	0.900	0.79
DeepXplain w/o Conf(e_t)	0.910	0.74
DeepXplain (full)	**0.915**	**0.86**

nation focuses on a smaller, more relevant subset of nodes and edges. DeepXplain produces more compact explanations (0.31 vs. 0.46, −32.6%), demonstrating its ability to localize critical attack patterns while filtering out noisy dependencies. In addition, we evaluate fidelity, which measures how well the generated explanations reflect the true decision-making behavior of the model. Fidelity is assessed by examining how much the model's prediction confidence degrades when the identified explanatory components are removed. DeepXplain achieves a higher fidelity score (0.79 vs. 0.62, +27.4%), indicating that its explanations capture the actual causal factors driving the model's predictions, rather than superficial correlations.

These results highlight a fundamental limitation of post-hoc explanation methods. While they provide interpretability after training, they do not influence the learning process and therefore cannot prevent reliance on spurious correlations. In contrast, DeepXplain integrates explanation signals directly into policy optimization, enforcing consistency between policy attention and graph-derived evidence. This results in explanations that are more concise, reliable, and causally grounded, leading to decisions that are both interpretable and operationally meaningful.

4) Ablation Study: The ablation results in Table II quantify the contribution of each component in the proposed XAI-guided objective. Removing the alignment loss $\mathcal{L}_{align}$ reduces the average stage-F1 from 0.915 to 0.900 (−1.6%) and decreases explanation confidence from 0.86 to 0.79 (−8.1%). This indicates that aligning policy attribution with graph-based evidence is essential for ensuring that the agent focuses on causally relevant features, thereby improving both robustness and interpretability. On the other hand, removing the confidence-based reward Conf(e_t) results in a smaller drop in stage-F1 (0.915 to 0.910, −0.5%) but a more significant reduction in explanation confidence (0.86 to 0.74, −14.0%). This suggests that the confidence term primarily enhances the reliability and consistency of explanations by encouraging the agent to favor actions supported by concentrated and high-quality evidence across different attack scenarios.

Overall, the full DeepXplain model achieves the best trade-off between defense effectiveness and explanation quality. The ablation study demonstrates that the alignment loss improves decision accuracy by enforcing evidence consistency, while the confidence reward enhances explanation reliability. Their combination enables DeepXplain to jointly optimize performance and interpretability, rather than treating explanation as a purely post-hoc artifact.

V. CONCLUSION

This paper presented *DeepXplain*, an XAI-guided extension of DeepStage for autonomous defense against multi-stage APT campaigns. The framework integrates provenance-based graph reasoning, temporal stage inference, and deep reinforcement learning with a unified explanation pipeline. Unlike post-hoc approaches, DeepXplain incorporates explanation signals directly into policy optimization through evidence alignment and confidence-aware reward shaping. Experimental results demonstrate improvements in both defense effectiveness and interpretability, achieving higher stage-wise performance and more reliable explanations. Future work will explore human-in-the-loop security operations and the integration of large language models for enhanced decision support.

REFERENCES

[1] T. V. Phan *et al.*, "Deepstage: Learning autonomous defense policies against multi-stage apt campaigns," in *IEEE International Conference on Cyber Security and Resilience (CSR)*, 2026. [Submitted]. Available on https://arxiv.org/abs/2603.16969.

[2] A. Alshamrani *et al.*, "A survey on advanced persistent threats: Techniques, solutions, challenges, and research opportunities," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 2, pp. 1851–1877, 2019.

[3] S. Milani, N. Topin, M. Veloso, and F. Fang, "Explainable reinforcement learning: A survey and comparative review," *ACM Computing Surveys*, vol. 56, no. 7, 2024.

[4] R. Ying *et al.*, "Gnnexplainer: generating explanations for graph neural networks," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

[5] T. M. Corporation, "MITRE Caldera: Automated Adversary Emulation Platform," 2024. Open-source adversary-emulation system used for breach-and-attack simulation of APT playbooks.

[6] X. Li *et al.*, "Can graph neural networks be adequately explained? a survey," *ACM Computing Surveys*, 2025.

[7] E. Dai *et al.*, "A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability," *International Journal of Automation and Computing*, 2024.

[8] A. T. Le *et al.*, "Automated apt defense using reinforcement learning and attack graph risk-based situation awareness," in *Proceedings of the Workshop on Autonomous Cybersecurity*, 2024.

[9] S. Thompson *et al.*, "Entity-based reinforcement learning for autonomous cyber defence," in *Proceedings of the Workshop on Autonomous Cybersecurity*, p. 56–67, 2024.

[10] M. U. Rehman, H. Ahmadi, and W. U. Hassan, "Flash: A comprehensive approach to intrusion detection via provenance graph representation learning," in *IEEE Symposium on Security and Privacy*, 2024.

[11] T. Bilot *et al.*, "Sometimes simpler is better: A comprehensive analysis of state-of-the-art provenance-based intrusion detection systems," in *USENIX Security Symposium*, 2025.

[12] A. T. Le *et al.*, "Automated apt defense using reinforcement learning and attack graph risk-based situation awareness," in *Proceedings of the Workshop on Autonomous Cybersecurity*, AutonomousCyber '24, (New York, NY, USA), p. 23–33, Association for Computing Machinery, 2024.