

APT-CGLP: Advanced Persistent Threat Hunting via Contrastive Graph-Language Pre-Training

Xuebo Qiu

College of Computer Science and
Technology, Zhejiang University of
Technology
Hangzhou, China
xueboqiu@zjut.edu.cn

Mingqi Lv*

College of Geoinformatics, Zhejiang
University of Technology
Zhejiang Key Laboratory of Visual
Information Intelligent Processing
Hangzhou, China
mingqilv@zjut.edu.cn

Yimei Zhang

College of Computer Science and
Technology, Zhejiang University of
Technology
Hangzhou, China
yimeizhang0229@zjut.edu.cn

Tieming Chen

College of Geoinformatics, Zhejiang
University of Technology
Zhejiang Key Laboratory of Visual
Information Intelligent Processing
Hangzhou, China
tmchen@zjut.edu.cn

Tiantian Zhu

College of Computer Science and
Technology, Zhejiang University of
Technology
Hangzhou, China
ttzhu@zjut.edu.cn

Qijie Song

College of Geoinformatics, Zhejiang
University of Technology
Hangzhou, China
songqijie@zjut.edu.cn

Shouling Ji

College of Computer Science and
Technology, Zhejiang University
Hangzhou, China
sji@zju.edu.cn

Abstract

Provenance-based threat hunting identifies Advanced Persistent Threats (APTs) on endpoints by correlating attack patterns described in Cyber Threat Intelligence (CTI) with provenance graphs derived from system audit logs. A fundamental challenge in this paradigm lies in the *modality gap*—the structural and semantic disconnect between provenance graphs and CTI reports. Prior work addresses this by framing threat hunting as a graph matching task: 1) extracting attack graphs from CTI reports, and 2) aligning them with provenance graphs. However, this pipeline incurs severe *information loss* during graph extraction and demands intensive manual curation, undermining scalability and effectiveness.

In this paper, we present **APT-CGLP**, a novel cross-modal APT hunting system via Contrastive Graph-Language Pre-training, facilitating end-to-end semantic matching between provenance graphs and CTI reports without human intervention. First, empowered by the Large Language Model (LLM), APT-CGLP mitigates data scarcity by synthesizing high-fidelity provenance graph-CTI report pairs, while simultaneously distilling actionable insights from noisy web-sourced CTIs to improve their operational utility. Second, APT-CGLP incorporates a tailored multi-objective training

algorithm that synergizes contrastive learning with inter-modal masked modeling, promoting cross-modal attack semantic alignment at both coarse- and fine-grained levels. Extensive experiments on four real-world APT datasets demonstrate that APT-CGLP consistently outperforms state-of-the-art threat hunting baselines in terms of accuracy and efficiency.

CCS Concepts

• **Security and privacy** → **Intrusion detection systems.**

Keywords

Advanced Persistent Threat, Threat Hunting, Provenance Graph, Multimodal Learning

ACM Reference Format:

Xuebo Qiu, Mingqi Lv, Yimei Zhang, Tieming Chen, Tiantian Zhu, Qijie Song, and Shouling Ji. 2026. APT-CGLP: Advanced Persistent Threat Hunting via Contrastive Graph-Language Pre-Training. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '26)*, August 09–13, 2026, Jeju Island, Republic of Korea. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3770854.3780275>

1 Introduction

Advanced Persistent Threats (APTs) pose a significant risk due to their persistence, stealth, and potentially severe impact [38]. To combat these threats, security practitioners engage in the collection, analysis, and interpretation of Cyber Threat Intelligence (CTI) that offers actionable insights into offensive tactics, techniques, and procedures (TTPs) [47]. Following this foundation, *threat hunting* has

*Corresponding author.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

KDD '26, Jeju Island, Republic of Korea

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2258-5/2026/08

<https://doi.org/10.1145/3770854.3780275>

emerged as a proactive defense strategy that utilizes CTI-recorded attack knowledge to identify system compromises [41].

Given the complexity of APT attacks, recent studies [5, 6, 16, 40] adopt *data provenance* to convert audit logs into provenance graphs (see Figure 1c) that capture structured interactions (e.g., file writing) among system entities (i.e., process, file, socket), enabling comprehensive system behavior analysis. Capitalizing on this, threat hunting is typically formulated as two categories of matching tasks. The first, *signature-based matching*, searches for CTI-documented Indicators of Compromise (IoCs) within provenance graphs. While efficient, it is highly vulnerable to evasion techniques like IoC obfuscation [18]. The second, *behavior pattern-based matching*, aims to mine abstract attack patterns from CTI reports and match them with provenance graphs [5, 15, 42]. This approach offers greater generalization and has seen broad adoption in practice [6, 33, 34, 37]. Concretely, it adheres to a two-stage process: 1) **Query Graph Construction**, which derives attack (query) graphs from CTI reports—either manually or via automated extraction framework—to bridge the natural language modality of CTI reports with the graph modality of provenance data; and 2) **Graph Pattern Alignment**, which applies heuristic rules or graph learning algorithms to match query graphs against suspicious provenance graphs, aiming to uncover behavior patterns indicative of APT attacks. Despite their potential, they suffer from a major limitation: *transforming unstructured CTI reports into query graphs incurs substantial information loss* [5, 6, 37], primarily due to the complexity and ambiguity of attack narratives (see Section 2). As a result, extensive human intervention is required to recover missing semantics from the query graph, which undermines the efficacy and scalability of such methods.

Meanwhile, multimodal learning has made remarkable strides across various domains [31, 32, 39, 56, 60], with growing emphasis on integrating information from heterogeneous sources to improve cross-modal task performance. A prominent milestone, CLIP [39], attains strong zero-shot capabilities on multiple benchmarks by learning unified representations from large-scale image-caption pairs. Such advances inspire a pivotal question: *can APT hunting be transformed into a similar end-to-end matching process between provenance graphs and CTI reports that share consistent attack semantics, thereby circumventing the reliance on manual CTI engineering?* However, unlike image-caption alignment, bridging gaps between provenance graphs and CTI reports introduces unique challenges:

C1. Significant Modality Gaps. Provenance graphs provide low-level, structured representations of system behaviors from audit logs, while CTI reports contain high-level APT narratives in unstructured natural language. Aligning these disparate semantic abstraction and structural formats remains a major challenge.

C2. Lack of Cross-Modal Supervision. Although CTI reports are widely available, matched provenance graphs remain extremely scarce due to the difficulty of acquiring authentic APT data. Meanwhile, the prohibitive cost of APT simulations (e.g., the DARPA TC program [7]) further limits the manual curation of training samples, forming a critical bottleneck for multimodal learning.

C3. Noisy and Verbose CTIs. Unlike the concise and semantically focused image captions in vision-language tasks, CTI reports are often verbose (averaging over 3,000 words [14]) and laden with noisy content (e.g., advertisements), which impedes the effective exploitation of inherent attack knowledge for threat hunting.

To address these challenges, we present **APT-CGLP**, a pioneering cross-modal APT hunting system via Contrastive Graph-Language Pre-training, enabling end-to-end threat hunting without manual intervention. For challenge C1, we design a multi-objective training algorithm that integrates contrastive learning and inter-modal masked modeling to comprehensively align attack semantics between provenance graphs and CTI reports. More specifically, contrastive learning first aligns their global semantic representations, while a cross-attention mechanism operates over their node- and token-level embeddings in conjunction with masked modeling, thereby enforcing fine-grained semantic consistency. For challenge C2, we propose a Graph2CTI module that leverages Large Language Models (LLMs) to automatically synthesize CTI-style reports from attack-free provenance graphs to augment training data. For challenge C3, we introduce a CTI denoising module powered by the chain-of-thought reasoning capacity of LLMs to distill concise and actionable threat insights from noisy CTI reports. Extensive experiments on four real-world APT datasets [1, 2] demonstrate that APT-CGLP outperforms state-of-the-art threat hunting baselines, while eliminating the need for manual query graph engineering.

In summary, the contributions of this work are as follows:

- We propose APT-CGLP, an end-to-end APT hunting system that facilitates threat hunting from provenance graphs and CTI reports without requiring human intervention.
- We design a multi-objective training algorithm that combines contrastive learning with inter-modal masked modeling to enhance semantic alignment between provenance graphs and CTI reports at both coarse- and fine-grained levels.
- We design a series of LLM-driven optimization strategies to improve generalization, including a Graph2CTI module for synthesizing provenance graph-CTI report training pairs, and a CTI denoising module for streamlining real-world CTI reports.
- We conduct extensive experiments on real-world APT attack datasets, demonstrating that APT-CGLP outperforms state-of-the-art methods in both accuracy and efficiency.

2 Related Work

Provenance-based Threat Hunting. Current approaches [5, 6, 37, 51] commonly formulate threat hunting as a graph pattern matching task. They first transform CTI-recorded attack patterns into structured query graphs, and then apply pattern matching algorithms to align them with provenance graphs for APT hunting.

For instance, Poirot [37] converts IoC interactions into query graphs, and matches against provenance graphs using a cost-aware similarity function that prioritizes causal relationships and information flows. Threatraptor [24] introduces a custom natural language processing (NLP) pipeline for CTI knowledge extraction and a domain-specific threat query language for large-scale audit log analysis. Graph learning-based methods like MEGR-APT [6], DeepHunter [51] extract query graphs through manual CTI analysis, and convert the matching problem into vector similarity evaluation through neural representations. ProvG-Searcher [5] utilizes the order embedding technique to learn subgraph entailment relationships, enabling efficient subgraph matching via online comparisons using precomputed provenance graph representations.

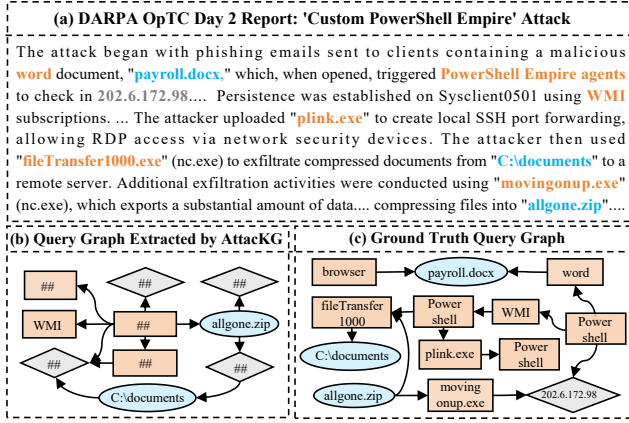


Figure 1: A motivating example, where subfigure (a) denotes the denoised CTI report from the DARPA OpTC engagement, and subfigures (b) and (c) represent the query graphs generated by AttackKG (## denotes unknown entities) and the ground truth, respectively.

Despite their successes, these methods are fundamentally constrained by their *heavy reliance on labor-intensive query graph construction*. Although NLP techniques have been introduced to automate this transformation, they remain error-prone. Consider a denoised CTI report in Figure 1a, which describes an APT campaign from the OpTC engagement [2] involving PowerShell Empire for data exfiltration. Nonetheless, even after applying our CTI denoising module (see §4.3) to refine the report and improve its downstream usability, AttackKG [33], a state-of-the-art CTI extraction framework, still produces a query graph (Figure 1b) that captures only 30% of the key system entities compared to the ground truth (Figure 1c), severely limiting its effectiveness for threat hunting.

Multimodal Learning. Multimodal learning aims to build models that can jointly process and align information from heterogeneous modalities by learning a shared embedding space [59, 62]. CLIP [39] is a representative work that aligns visual and textual modalities using a simple yet effective contrastive learning approach, achieving strong zero-shot performance across various tasks. Follow-up work has extended this paradigm with more sophisticated alignment mechanisms [30–32, 49, 55]. For example, BLIP [31] introduces a bootstrapping approach that combines contrastive learning with captioning and filtering objectives to better handle noisy web data, while BLIP-2 [30] further improves efficiency by incorporating a lightweight querying transformer in a two-stage training framework. Beyond vision-language domains, recent studies have also explored fusing molecular graphs with textual descriptions for molecular retrieval [20, 21, 44]. Nonetheless, multimodal learning remains underexplored in the context of threat hunting, primarily due to the scarcity of training data and the alignment challenges between provenance graphs and CTI reports, as discussed earlier.

3 Preliminaries

3.1 Definitions

Provenance Graph. A provenance graph provides a structured representation of audit logs, enabling comprehensive analysis of

system behaviors. Formally, it is defined as $G = (V, E)$, where V represents the set of nodes corresponding to system entities (i.e., process, socket, file), and E is the set of directed edges indicating system events. Each edge $e \in E$ is defined as the tuple $e = \langle s, a, o, t \rangle$, where s is a subject entity (a process) initiating the action a (e.g., write), $o \in V$ is the object entity, and t is the timestamp of the event. **Cyber Threat Intelligence (CTI).** CTI refers to evidence-based security knowledge, including IoCs, TTPs, and threat assessments (Figure 1a), which supports informed defense decisions [47]. CTI reports consolidate threat data from various sources (e.g., audit logs, vulnerability databases, threat intelligence feeds) and are further enriched through manual examination and expert interpretation. Such insights empower organizations to proactively anticipate, identify, and respond to cyber threats more effectively.

3.2 Threat Model

Our threat model assumes a trusted computing base encompassing the operating system, hardware, and audit frameworks, in line with prior work [6, 13, 16, 28, 61]. Potential threats such as audit log tampering or hardware-level exploits are deemed beyond the scope of this work. Furthermore, the CTI reports utilized for threat hunting are regarded to be accurate, trustworthy, and free from adversarial manipulation. While adversaries may adapt the implementation details of attack techniques to evade detection, we posit that the core semantics and objectives of these attacks remain unchanged.

4 System Architecture

This section outlines the APT-CGLP architecture, as illustrated in Figure 2. We first introduce the Graph2CTI module that generates high-quality provenance graph-CTI report pairs for cross-modal supervision (§4.1). Next, we describe the multi-objective training framework, designed to align cross-modal attack semantics across both global and fine-grained levels through complementary learning tasks (§4.2). Following this, the CTI denoising module is discussed, aimed at refining multi-sourced reports to enhance operational utility (§4.3). Finally, we detail the threat hunting module, which incorporates a two-stage retrieval strategy to balance detection precision and computational efficiency at scale (§4.4).

4.1 Graph2CTI Module

In contrast to the abundance of image-caption datasets, the threat hunting domain faces a notable scarcity of paired provenance graph-CTI report samples for training (Challenge C2). We tackle this gap through two key insights. First, audit logs provide a holistic and fine-grained account of system activities, and have long served as a trusted basis for forensic analysis [22, 26, 61], making them a reliable data source for reverse-generating CTI reports. Second, unlike intrusion detection that aims to explicitly identify attack patterns, threat hunting emphasizes semantic alignment between behavior patterns described in CTI reports and those embedded in provenance graphs [5, 6]. This fundamental shift allows us to exploit the ubiquitous attack-free audit logs to synthesize high-quality training pairs for cross-modal supervision. These observations motivate the design of our Graph2CTI module, which first samples semantically rich subgraphs from benign provenance graphs, and then converts

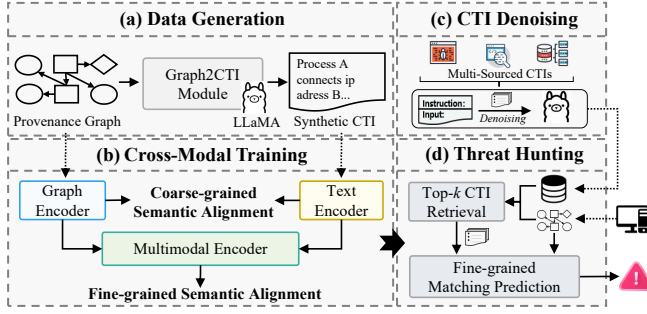


Figure 2: APT-CGLP architecture. (a) Synthetic provenance graph-CTI report pair generation for training; (b) Cross-modal semantic alignment via multi-objective pre-training; (c) CTI report denoising to enhance the usability; and (d) Two-stage retrieval using trained encoders in (b) to balance efficiency and precision in threat hunting.

them into structured interaction triplets that can be processed by LLMs to generate corresponding CTI reports.

Subgraph Sampling. For constructing paired training samples, we begin by sampling subgraphs representing different activity instances as references. While random sampling is a straightforward approach, it often yields skewed subgraphs dominated by monotonous interactions (e.g., repetitive file reads), which lack the behavior diversity inherent to real-world APT scenarios [16, 34] and thus hinder the model’s ability to learn generalizable features. To overcome this, we design a heuristic-driven sampling algorithm that prioritizes attack-relevant patterns. It performs depth-first traversal, guided by the following three heuristics:

- **Depth Limitation:** The traversal depth is restricted to 2-3 layers, corresponding to behavior paths of 5-7 steps—typical of realistic APT campaigns [6].
- **Size Constraint:** Each subgraph is limited to 10-20 nodes, a range derived from extensive statistical analysis of CTI reports.
- **Diverse Interactions:** Subgraphs are encouraged to include diverse entity types (e.g., processes, files, sockets) to emulate multi-stage attacks, where adversaries may initiate processes, establish remote connections, and exfiltrate files [53].

These heuristics ensure that sampled subgraphs retain behavior semantics resembling APTs, enabling the generation of informative training pairs. We provide sampling pseudocode in Appendix A.

Synthetic CTI Generation. LLMs have demonstrated language understanding and generation capabilities on par with domain experts [25, 57]. We leverage this capability to translate sampled subgraphs into coherent CTI reports in an automated manner.

First, each subgraph is transformed into a set of interaction triplets, such as *(word.exe, read, payroll.docx)* in Figure 1, which can be readily processed by LLMs. Second, we employ in-context learning [10] to guide the LLM in compiling these triplets into CTI reports. Concretely, we craft a structured prompt containing three parts: 1) a **task description** that articulates the CTI generation objective, ensuring that the LLM comprehends the intended task; 2) an **in-context example** that demonstrates a specific mapping from triplets to the corresponding CTI report, enabling the LLM to internalize the transformation patterns; and 3) an **input section** that supplies the sequence of triplets to be translated. The full prompt

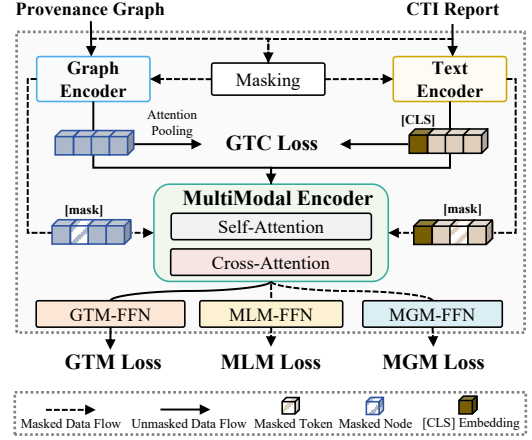


Figure 3: The training framework of APT-CGLP. GTC: graph-text contrastive, GTM: graph-text matching, MGM: masked graph modeling, MLM: masked language modeling, FFN: feed forward network.

is subsequently submitted to LLMs, which outputs a synthetic CTI report that captures the high-level attack narrative of the original subgraph. Prompt design details are included in Appendix B.

4.2 Cross-Modal Training Module

With ample paired training samples, an intuitive approach is to adopt cross-modal contrastive learning for semantic alignment, following the CLIP paradigm [39]. However, aligning provenance graphs with CTI reports presents domain-specific challenges. First, unlike CLIP that benefits from powerful visual models pretrained on large-scale images, no such models exist for provenance graphs. Second, CLIP focuses on coarse-grained global alignment [12], but APT campaigns often share similar high-level tactics (e.g., Collection, Exfiltration) [53], making global alignment insufficient for fine-grained pattern differentiation. Moreover, current graph-text contrastive methods [52, 56, 60] assume relatively homogeneous graphs and brief textual inputs, which are misaligned with our setting—where provenance graphs encode low-level, heterogeneous system events, and CTI reports contain high-level, unstructured behavior narratives (Challenge C1). This semantic and structural asymmetry poses significant obstacles for cross-modal training.

To tackle these issues, we propose a training framework that bridges provenance graphs and CTI reports across both coarse- and fine-grained semantic levels. As depicted in Figure 3, the framework comprises three major components—a graph encoder, a text encoder, and a multimodal encoder—that are jointly optimized using a set of tailored objectives and task-specific heads to achieve cross-modal semantic alignment.

4.2.1 Encoders. APT-CGLP integrates two unimodal encoders to independently process provenance graphs and CTI reports, along with a transformer-based multimodal encoder to learn their implicit correlations and produce unified representations.

Graph Encoder. To encode the structural interaction semantics of provenance graphs, we employ the Graph Isomorphism Network

Table 1: Node attributes in provenance graphs.

Node Type	Attribute	Example Instance
Process	Type: CommandLine	process: movingonup.exe fun.com 81
File	Type: Filepath	file: /142.20.61.135/share
Socket	Type: IP Address:Port	socket: 127.0.0.1:0

(GIN) [54], denoted as f_G , for its powerful topological discrimination capacity. Node attributes (see Table 1) and edge types are initialized as graph features via our text encoder. Each provenance graph G_i is then processed by f_G , which aggregates multi-hop neighborhoods for each node to encode its contextual behaviors (detailed in Appendix D). The resulting node embeddings are defined as:

$$\mathbf{H}_{G_i} = \{\mathbf{h}_{n_1}, \mathbf{h}_{n_2}, \dots, \mathbf{h}_{n_{|V_i|}}\} = f_G(G_i; \theta_G), \quad (1)$$

where $\mathbf{h}_{n_j} \in \mathbb{R}^d$ is the j -th node embedding, $|V_i|$ is the number of nodes in G_i , and θ_G denotes the parameters of GIN.

Text Encoder. We use BERT [19] as the text encoder f_T , due to its effectiveness in capturing bidirectional causal dependencies among system entities within CTI reports. Each CTI report T_i is first tokenized, and then passed through self-attention layers to produce token-level contextual embeddings:

$$\mathbf{H}_{T_i} = \{\mathbf{h}_{t_{[\text{CLS}]}}^{\text{CLS}}, \mathbf{h}_{t_1}, \mathbf{h}_{t_2}, \dots, \mathbf{h}_{t_{|T_i|}}\} = f_T(T_i; \theta_T), \quad (2)$$

where $\mathbf{h}_{t_j} \in \mathbb{R}^d$ is the embedding for token t_j , and $\mathbf{h}_{t_{[\text{CLS}]}}^{\text{CLS}}$ denotes the global semantic representation of the CTI report [19, 32]. The parameters of BERT are denoted by θ_T .

Multimodal Encoder. To learn semantic associations between provenance graphs and CTI reports, we introduce a multimodal encoder f_M , built on a transformer backbone. The encoder first applies self-attention over \mathbf{H}_{T_i} to capture intra-text dependencies, and subsequently utilizes cross-attention mechanisms to integrate structured node semantics \mathbf{H}_{G_i} . Such a design enables the model to learn detailed token-node interactions and achieve precise semantic alignment across modalities. The final multimodal representation is computed as:

$$\mathbf{z}_m = f_M(\mathbf{H}_{G_i}, \mathbf{H}_{T_i}; \theta_M), \quad (3)$$

where θ_M denotes the parameters of f_M , and \mathbf{z}_m is the resulting joint representation for downstream alignment and threat hunting.

4.2.2 Training Strategy. Global semantic alignment enforces coarse-grained consistency across modalities but lacks the sensitivity to differentiate attack campaigns with similar patterns. We mitigate this by adopting a hierarchical learning strategy: the global alignment objective establishes a fundamental cross-modal semantic coherence, while a set of auxiliary tasks offers fine-grained supervision to resolve subtle semantic inconsistencies.

Graph-Text Contrastive Learning (GTC). This objective enforces global semantic alignment between provenance graphs and CTI reports by maximizing agreement between matched pairs while pushing apart mismatched ones in the shared embedding space.

Specifically, given a provenance graph-CTI report pair (G_i, T_j) , the unimodal encoders f_G and f_T are used to extract their respective overall representations. The graph representation is derived via an attention pooling mechanism over node embeddings: $\mathbf{z}_{G_i} = \sum_{n \in V_i} \alpha_n \mathbf{h}_n$, where α_n is the attention weight of node n . The CTI representation \mathbf{z}_{T_j} is extracted as the hidden state of its [CLS] token:

$\mathbf{z}_{T_j} = \mathbf{h}_{t_{[\text{CLS}]}}$. We then compute the softmax-normalized similarity between the paired representations across a mini-batch as:

$$p_{ij}^{g2t}(G, T) = \frac{\exp(\text{sim}(\mathbf{z}_{G_i}, \mathbf{z}_{T_j})/\tau)}{\sum_{k=1}^B \exp(\text{sim}(\mathbf{z}_{G_i}, \mathbf{z}_{T_k})/\tau)}, \quad (4)$$

where $\text{sim}(\cdot, \cdot)$ denotes dot-product similarity, τ is a learnable temperature parameter, and B is the batch size. Using one-hot supervision $\mathbf{y}^{g2t} \in \{0, 1\}^{B \times B}$, where $y_{ij}^{g2t} = 1$ indicates a positive pair, we optimize the model with a cross-entropy loss:

$$\mathcal{H}(\mathbf{y}^{g2t}, \mathbf{p}^{g2t}(G, T)) = - \sum_{i=1}^B \sum_{j=1}^B y_{ij}^{g2t} \log(p_{ij}^{g2t}(G, T)), \quad (5)$$

$$\mathcal{L}_{g2t} = \mathbb{E}_{\{G, T\} \sim \mathcal{D}} \mathcal{H}(\mathbf{y}^{g2t}, \mathbf{p}^{g2t}(G, T)), \quad (6)$$

where $\{G, T\} \sim \mathcal{D}$ denotes a mini-batch sampled from the dataset \mathcal{D} , and \mathcal{L}_{g2t} is the graph-to-text contrastive loss. To enforce bidirectional semantic agreement, we compute a symmetric loss \mathcal{L}_{t2g} for text-to-graph matching, and define the total GTC loss as: $\mathcal{L}_{gtc} = (\mathcal{L}_{g2t} + \mathcal{L}_{t2g})/2$.

Graph-Text Matching (GTM). To capture fine-grained semantic alignment, we introduce a binary matching task that predicts whether the behavior semantics of a given pair are consistent. The multimodal encoder f_M produces a joint embedding \mathbf{z}_m that captures detailed interactions between graph nodes and CTI tokens. The \mathbf{z}_m is then passed through a feedforward network (GTM-FFN) to compute the matching probability. The GTM loss is defined as:

$$\mathcal{L}_{gtm} = \mathbb{E}_{\{G, T\} \sim \mathcal{D}} \mathcal{H}(\mathbf{y}^{gtm}, \mathbf{p}^{gtm}(G, T)), \quad (7)$$

where \mathbf{y}^{gtm} denotes ground-truth labels. To enhance learning efficiency, a *hard negative sampling* strategy [32] is employed to select non-matching pairs with high global similarities as challenging negatives, which sharpens the model's discriminative capacity.

Masked Language Modeling (MLM). This objective strengthens cross-modal understanding by training the model to recover masked tokens in a corrupted CTI report \hat{T} using both the visible text and the associated graph context. To construct \hat{T} , we randomly replace a subset of tokens in the original CTI report with the special token [MASK]. Then, we feed (\hat{T}, G) into f_M to integrate structural and textual information via cross-attention, and a task-specific head (MLM-FFN) outputs the vocabulary distribution probability $\mathbf{p}^{\text{mlm}}(G, \hat{T})$ of masked tokens. Let \mathbf{y}^{mlm} be the corresponding one-hot encoded ground-truth labels. The MLM loss is formulated as:

$$\mathcal{L}_{\text{mlm}} = \mathbb{E}_{\{G, \hat{T}\} \sim \mathcal{D}} \mathcal{H}(\mathbf{y}^{\text{mlm}}, \mathbf{p}^{\text{mlm}}(G, \hat{T})). \quad (8)$$

Masked Graph Modeling (MGM). Following the same principle as MLM, this task requires the model to reconstruct masked nodes in the provenance graph \hat{G} based on the unmasked nodes and the paired CTI report. Specifically, we randomly mask nodes in the original provenance graph to construct \hat{G} . Then, f_M applies cross-attention between nodes in \hat{G} and tokens in the paired CTI report T , followed by a prediction head (MGM-FFN) to predict masked node embeddings $\mathbf{p}^{\text{mgm}}(\hat{G}, T)$. Let \mathbf{y}^{mgm} represent the corresponding original node features. The training objective minimizes the mean squared error between the predicted and original node

embeddings:

$$\mathcal{L}_{\text{mgm}} = \mathbb{E}_{\{\hat{G}, T\} \sim \mathcal{D}} \left(\frac{1}{B} \sum_{i=1}^B \|y_i^{\text{mgm}} - p_i^{\text{mgm}}(\hat{G}, T)\|_2^2 \right), \quad (9)$$

The final loss function integrates above objectives through a weighted combination as follows:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{gic}} + (1-\alpha)(\mathcal{L}_{\text{gtm}} + \mathcal{L}_{\text{mlm}} + \mathcal{L}_{\text{mgm}}), \quad (10)$$

where $\alpha \in (0, 1)$ controls the trade-off between coarse- and fine-grained semantic alignment. A larger weight is placed on GTC to establish global semantic consistency, while the auxiliary tasks foster more nuanced cross-modal understanding.

4.3 CTI Denoising Module

Rather than being clean and structured like the synthetic CTI reports used for training, real-world CTI reports often contain substantial noise, such as website metadata [33], which obscures the core attack insights essential for threat hunting (Challenge C3). To mitigate this, we leverage the advanced language summarization capabilities of LLMs [57] to isolate and restructure actionable intelligence from raw CTI inputs, effectively filtering out extraneous content [58]. Particularly, we adopt a Chain-of-Thought (CoT) reasoning framework that decomposes the denoising process into logically ordered subtasks, emulating the analytical workflow of threat analysts. The CoT prompt comprises three reasoning stages:

- (1) **Entity Identification:** Scan the CTI report to extract core system entities (e.g., malicious payloads) involved in the attack.
- (2) **Interaction Extraction:** Infer both explicit and implicit interactions among entities based on contextual cues.
- (3) **Knowledge Distillation:** Consolidate the extracted interactions into a concise and temporally ordered attack narrative.

By guiding the LLM through this structured reasoning process, we obtain distilled intelligence that retains essential attack patterns while suppressing noise. Complete prompt templates are provided in Appendix C.

4.4 Threat Hunting Module

Effective threat hunting is enabled by integrating the trained multi-modal encoder and GTM-FFN to recursively estimate the matching probabilities between provenance graphs and CTI reports (see Section 4.2.2). However, directly applying this matching process over large-scale provenance graphs and an expanding CTI corpus is computationally prohibitive. To address this scalability bottleneck, we reformulate the threat hunting as a *two-stage retrieval pipeline*. The first stage narrows the search space via a lightweight similarity-based retrieval, while the second stage conducts fine-grained semantic matching for high-precision threat identification.

- (1) **Coarse-Grained Retrieval:** All CTI reports are embedded into dense representations $\{z_t^1, z_t^2, \dots, z_t^n\}$ using the text encoder and indexed in a vector database. Given a target provenance graph, the graph encoder computes its global embedding z_g . Top- k CTI candidates are then retrieved based on cosine similarity between z_g and stored CTI embeddings.
- (2) **Fine-Grained Matching:** The retrieved CTI candidates and z_g are passed into the multimodal encoder to produce joint representations $\{z_m^1, z_m^2, \dots, z_m^k\}$. These embeddings are then

fed into the GTM-FFN, which computes matching probabilities, enabling precise identification of the most relevant CTI report.

5 Evaluation

We conduct comprehensive experiments to investigate the following research questions (RQs):

- **RQ1:** How does APT-CGLP compare to state-of-the-art threat hunting systems?
- **RQ2:** What are the individual contributions of its core modules, namely training data augmentation, cross-modal training, and CTI denoising to overall performance?
- **RQ3:** How well does the two-stage retrieval strategy balance threat hunting precision and computational efficiency?
- **RQ4:** How effective is APT-CGLP in supporting alert validation?

5.1 Experiment Settings

This section describes the datasets, evaluation protocol, metrics, and the implementation and hardware configuration used in our experiments.

5.1.1 Datasets. We evaluate APT-CGLP on the DARPA TC E3 [1] and OpTC [2] datasets, both of which simulate real-world enterprise-scale APT campaigns. The E3 dataset spans two weeks of system audit logs, with the first week recording benign activity and the second week containing various attack scenarios (e.g., Nginx exploits, Firefox backdoors). We select the Cadets, Trace, and Theia subsets for evaluations due to their comprehensive audit coverage and well-documented reports. The OpTC dataset is the latest iteration of the DARPA TC project, offering over 17 billion events across 1,000 Windows hosts. It similarly comprises multiple days of benign activities followed by three days of sophisticated attack scenarios, such as “PowerShell Empire” and “Malicious Upgrade”. Audit logs generated from the victim hosts are utilized for evaluation.

5.1.2 Evaluation Protocol. We partition datasets into training and testing subsets, then establish comprehensive evaluation metrics.

Pre-training Strategy. We use only benign audit logs from the first week of each dataset for training. The Graph2CTI module transforms daily logs into provenance graphs and samples representative subgraphs, which are then processed by the LLM to generate synthetic CTI reports (Section 4.1). As summarized in Table 2, this process yields 45,225 provenance graph-CTI report pairs. Crucially, attack-related data is excluded during training data curation to prevent *data leakage* during evaluation [8].

Testing Strategy. For rigorous assessment, we curate a hybrid test corpus containing both malicious and benign provenance graphs, along with a large collection of web-sourced CTI reports.

- **Provenance Graphs.** We evaluate two types of subgraphs: *malicious subgraphs* for attack detection and *benign subgraphs* for assessing false alert filtering. All subgraphs are generated using the MEGR-APT sampling strategy [6]. First, anomalous entities are selected as seed nodes based on predefined heuristics. For malicious cases, seeds are entities with names matching IoCs and interaction timestamps aligning with attack windows recorded in DARPA reports. This cross-validation guarantees accurate annotation. For benign cases, seeds are derived from: 1) entities in *benign logs* matching IoC names, and 2) false alarms reported by

Table 2: Training and testing datasets. PG: provenance graph.

Dataset	Pre-training		Testing		
	# PG	# CTIs	# PG	# CTIs	
				# DARPA	# Web-Sourced
E3-CADETS	7,524	721	4	5,172	
E3-THEIA	10,263	318	2		
E3-Trace	20,781	873	2		
OpTC	6,657	490	3		
Total	45,225	2,402		5,183	

the APT detection system Magic [28]. Anchored on each seed node, subgraphs are then constructed by expanding to include neighboring seeds and process nodes within a 2-3-hop radius.

- **CTI Corpus.** The CTI corpus composes two sources: official DARPA reports and web-sourced reports from security vendors (e.g., MITRE [17], VirusTotal [46], Microsoft [36]). To ensure quality and relevance, we apply regex-based filtering to an initial pool of 10,618 reports, retaining 5,172 CTI reports that contain detailed IoC and TTP descriptions. Keyword analysis confirms that these reports cover attack vectors such as phishing, living-off-the-land, and backdoor exploits. All CTI reports are processed through our denoising module to distill core insights, forming the external knowledge base for threat hunting.

Table 2 presents the final testing dataset, which consists of 2,402 subgraphs and 5,183 CTI reports. Among them, 2,094 subgraphs are generated by Magic and are used for alert validation evaluation, while the remaining subgraphs (with 1.2% malicious) are generated by MEGR-APT and used for threat hunting evaluation.

5.1.3 Evaluation Metrics. We define a positive match between a provenance graph and a CTI report when: 1) their unimodal embedding similarity exceeds threshold λ , and 2) GTM-FFN predicts a match based on multimodal embeddings. Otherwise, it is a negative match. True positives occur when malicious subgraphs match their paired CTI report *exclusively*, whereas false negatives indicate unmatched malicious subgraphs. True negatives represent benign subgraphs with no CTI matches, while false positives denote benign subgraphs incorrectly matched to CTI reports. Our evaluations use these definitions with standard metrics: Recall, Precision, Accuracy, False Positive Rate (FPR), and F1-Score.

5.1.4 Environment and Implementation. All experiments were conducted on an Ubuntu 20.04 server equipped with an Intel Xeon Gold 5128 CPU, 128 GB of RAM, and two NVIDIA RTX 4090 GPUs.

The implementation of APT-CLIP comprises ~5,000 lines of Python codes. We adopt the BERT-Base model [3] as a pre-trained text encoder and fine-tune the last two layers. The graph encoder is implemented as a three-layer GIN network with its dimensionality aligned to that of the text encoder. A custom transformer-like model is designed as the multimodal encoder for cross-modal integration. During training, the token masking ratio for the MLM objective is set to 15%, and one node was randomly masked for the MGM objective. Optimization is performed using AdamW with a learning rate of 2×10^{-4} and a weight decay of 0.01. A cosine annealing schedule is applied over 100 epochs, decaying the learning rate to a minimum of 1×10^{-5} . During the first 7 epochs, a warmup phase is employed. The loss-balancing weight is set to $\alpha = 0.7$. To mitigate overfitting, regularization and dropout strategies were implemented during model training. The vector database for storing

and retrieving CTI report embeddings is based on FAISS [29]. The optimal number of candidate CTI in the two-stage retrieval strategy was fixed at $k = 10$, and the similarity threshold was set to $\lambda = 0.5$. Finally, the open-source LLaMA3-8B model [4] is deployed locally for synthetic training data generation and CTI report denoising, providing lower computational overhead and improved security when processing sensitive intelligence data.

5.2 RQ1: Threat Hunting Performance

We compare APT-CGLP against three state-of-the-art baselines:

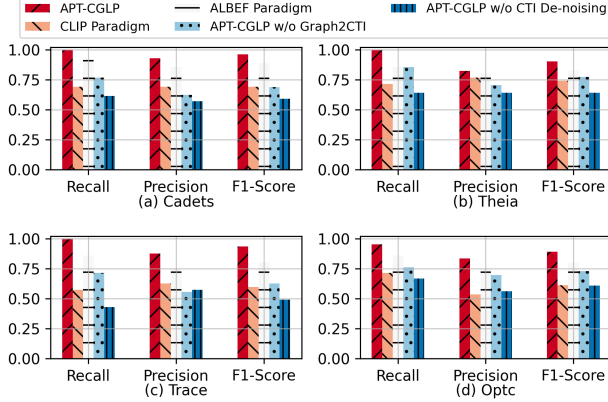
- **MEGR-APT** [6]: Learns attack graph representations for both query and provenance graphs using abstract node and edge features, followed by a Neural Tensor Network [43] for alignment.
- **ProvG-Searcher** [5]: Formulates threat hunting as a subgraph entailment problem using GNNs and order embeddings [27], with heuristic node abstractions for feature augmentation.
- **TREC** [34]: A heterogeneous graph attention network [50] encodes the behavior features of provenance subgraphs and CTI-derived query graphs into a shared embedding space via a siamese architecture, enabling few-shot recognition of similar patterns.

All baselines operate on CTI-derived query graphs for threat hunting. To ensure a rigorous comparison, we construct two variants for each: 1) the **w/o HI** variant utilizes query graphs automatically extracted from denoised CTI reports using AttackKG [33], simulating a fully automated pipeline without human involvement (HI); 2) the **w/ HI** variant uses manually refined query graphs with complete semantics, consistent with their original implementations and reflective of state-of-the-art performance. Evaluation results are presented in Table 3, with key findings summarized as follows:

- **Impact of Human Involvement.** The w/o HI variants suffer from clear performance drops. On the E3-Cadets dataset, TREC w/o HI achieves 0.581 F1-Score, whereas incorporating human-in-the-loop intervention boosts the score to 0.877—a notable improvement of 29.6%. Similarly, ProvG-Searcher jumps from 0.452 (w/o HI) to 0.788 (w/ HI). This trend is consistent across datasets, confirming that semantic gaps in auto-generated query graphs degrade detection capability. The performance disparity highlights current CTI parsing limitations and the continued need for human curation to preserve semantic fidelity.
- **Superiority of APT-CGLP.** APT-CGLP achieves consistently strong performance across all datasets, setting a new benchmark for fully automated systems. Notably, it surpasses the w/o HI baselines by over 30% in F1-Score. The Recall for APT-CGLP is also perfect across three datasets and maintains lower FPRs. These results demonstrate its ability to capture both holistic consistent attack semantics automatically and robustly, outperforming baselines even with handcrafted query graphs.
- **Comparison with Human-Tuned Baselines.** Remarkably, APT-CGLP matches or exceeds the performance of w/ HI baselines on most datasets. For instance, it improves F1-Score upon MEGR-APT (w/ HI) by 6.6% on E3-Cadets and by 5.8% on E3-Trace. The only dataset where it slightly lags is E3-Theia, likely due to ambiguous CTI descriptions that cause several matches with normal provenance graphs. However, its FPR remains low

Table 3: Comparisons of APT-CGLP with state-of-the-art threat hunting methods across DARPA datasets. Best results are bolded, and second-best are underlined. ↑: higher is better. ↓: lower is better. Rec.: Recall, Pre.: Precision, Acc.: Accuracy, F1.: F1-Score.

Dataset Model	E3-Cadets					E3-Theia					E3-Trace					OpTC				
	Rec. ↑	FPR ↓	Pre. ↑	Acc. ↑	F1. ↑	Rec. ↑	FPR ↓	Pre. ↑	Acc. ↑	F1. ↑	Rec. ↑	FPR ↓	Pre. ↑	Acc. ↑	F1. ↑	Rec. ↑	FPR ↓	Pre. ↑	Acc. ↑	F1. ↑
ProvG-Searcher w/o HI	0.538	0.212	0.389	0.738	0.452	0.643	0.226	0.429	0.746	0.514	0.429	0.219	0.115	0.759	0.182	0.381	0.264	0.222	0.677	0.281
MEGR-APT w/o HI	0.615	0.192	0.444	0.769	0.516	0.786	0.132	0.611	0.851	0.688	0.571	0.152	0.200	0.830	0.296	0.571	0.179	0.387	0.780	0.462
TREC w/o HI	0.692	0.173	0.500	0.799	0.581	0.714	0.151	0.556	0.821	0.625	0.714	0.157	0.233	0.835	0.351	0.524	0.170	0.380	0.780	0.441
ProvG-Searcher w/ HI	1.000	0.135	0.650	0.892	0.788	1.000	0.113	0.700	0.910	0.824	1.000	0.056	0.636	0.949	0.778	1.000	0.085	0.700	0.929	0.824
MEGR-APT w/ HI	1.000	0.058	0.813	0.954	0.897	1.000	0.038	0.875	0.970	0.933	1.000	0.028	0.778	0.974	0.875	1.000	0.057	0.778	0.953	0.875
TREC w/ HI	1.000	0.077	0.765	0.938	0.877	1.000	0.075	0.778	0.941	0.875	1.000	0.021	0.824	0.981	0.903	1.000	0.075	0.724	0.937	0.840
APT-CGLP	1.000	0.019	0.929	0.985	0.963	1.000	0.057	0.824	0.955	0.903	1.000	0.014	0.875	0.987	0.933	0.952	0.038	0.833	0.961	0.889

**Figure 4: Results of ablation studies across different datasets.**

(0.057), suggesting these errors are well-contained. Overall, APT-CGLP offers a high-performance, scalable alternative to manually curated systems, significantly lowering operational cost.

5.3 RQ2: Ablation Studies

We conduct ablation studies to evaluate the contributions of each component—training data augmentation, different cross-modal training strategies, and CTI denoising—on the overall threat hunting performance of APT-CGLP. Results are presented in Figure 4.

Effect of Training Data Augmentation. We assess the efficacy of Graph2CTI module by comparing it to a variant that uses naive random subgraph sampling (APT-CGLP w/o Graph2CTI) during training. As illustrated in Figure 4, removing Graph2CTI leads to over 10% performance drop, likely due to the lack of semantic diversity and behavior richness in randomly sampled graphs. This confirms that training with informative subgraphs—mirroring real-world attack patterns—is critical for model generalization.

Efficacy of Fine-Grained Semantic Alignment. We evaluate the effect of fine-grained alignment by comparing APT-CGLP with a CLIP-style variant that only performs global contrastive learning at the embedding space (the CLIP paradigm). This variant yields the second-lowest F1-Score across datasets, highlighting that global alignment alone is insufficient to bridge the structural and semantic modality gap. In contrast, our model’s use of a multimodal encoder with cross-attention achieves over 20% gains in F1-Score, validating the necessity of modeling fine-grained node-token interactions.

Role of Inter-Modal Masked Modeling. To evaluate the effectiveness of inter-modal masked modeling strategy, we compare

APT-CGLP with ALBEF [32], a prior framework that uses only masked language modeling for vision-language alignment (the ALBEF paradigm). Figure 4 indicates that APT-CGLP consistently outperforms ALBEF in F1-Scores. These findings underscore the essential role of bidirectional cross-modal supervision for effectively capturing subtle semantic cues critical to robust threat hunting.

Influence of CTI Denoising. We assess the utility of the CTI denoising module by comparing APT-CGLP to a variant using noisy CTI reports (APT-CGLP w/o CTI denoising). Figure 4 shows that denoising consistently improves Recall by over 30% across all datasets. This substantial improvement confirms that noisy content significantly hinders CTI utilization. By distilling concise, attack-relevant descriptions, both precision and robustness are improved.

5.4 RQ3: Practicality of Retrieval Strategy

Effectiveness Analysis. We assess the effectiveness of the proposed two-stage retrieval strategy by comparing it with a similarity-based baseline and examining the influence of the candidate pool size k . The baseline approach retrieves the CTI report with the highest similarity score as the threat hunting result. For APT-CGLP, we assess two configurations with $k = 10$ and $k = 20$. The experimental results in Table 4 yield the following observations:

- **Improved Precision via Two-Stage Retrieval.** APT-CGLP consistently outperforms the similarity-based baseline across all benchmarks. The performance gains stem from the second-stage refinement, which enables fine-grained semantic alignment and effectively reduces false matches with benign subgraphs. This hierarchical design boosts Recall and overall robustness.
- **Candidate Pool Size Trade-offs.** Increasing the candidate size k from 10 to 20 improves Recall (e.g., on the OpTC dataset) by broadening the search space. However, this also leads to a marginal increase in FPRs due to more semantically relevant candidates that are mistakenly matched. These findings suggest that k should be adjusted according to deployment requirements.

System Overhead. We further assess the computational overhead of APT-CGLP using the E3-Cadets dataset, focusing on memory usage and inference latency under varying candidate sizes $k \in \{10, 20, 30, 40, 50\}$ across both CPU and GPU environments.

- **Memory Cost.** The total memory cost consists of (1) offline storage for CTI embeddings within a vector database (e.g., FAISS), and (2) runtime memory for the two-stage retrieval pipeline. Offline storage scales linearly with the corpus size, requiring approximately 0.5 MB per 1,000 CTI embeddings. As shown in Figure 5a, runtime memory is dominated by model weights and external dependencies, remaining stable at approximately ~1 GB

Table 4: Comparison of similarity-based retrieval and two-stage retrieval strategies in threat hunting. Retr.: retrieval.

Dataset	Strategy	Rec. ↑	FPR ↓	Pre. ↑	Acc. ↑	F1 ↑
E3-Cadets	Similarity-based Retr.	0.923	0.058	0.800	0.938	0.857
	Two-stage Retr. ($k@10$)	1.000	0.019	0.929	0.985	0.963
	Two-stage Retr. ($k@20$)	1.000	0.038	0.867	0.969	0.929
E3-Theia	Similarity-based Retr.	0.929	0.075	0.765	0.925	0.839
	Two-stage Retr. ($k@10$)	1.000	0.057	0.824	0.955	0.904
	Two-stage Retr. ($k@20$)	1.000	0.094	0.737	0.925	0.849
E3-Trace	Similarity-based Retr.	1.000	0.042	0.700	0.962	0.824
	Two-stage Retr. ($k@10$)	1.000	0.014	0.875	0.987	0.933
	Two-stage Retr. ($k@20$)	1.000	0.028	0.778	0.974	0.875
OpTC	Similarity-based Retr.	0.905	0.094	0.655	0.906	0.760
	Two-stage Retr. ($k@10$)	0.952	0.038	0.833	0.961	0.889
	Two-stage Retr. ($k@20$)	1.000	0.066	0.750	0.945	0.857

Table 5: The performance gains generated by APT-CGLP through alert validation.

Gains	E3-Cadets	E3-Theia	E3-Trace	OpTC
AFR ↑	0.981	0.943	0.986	0.962
TRR ↑	1.000	1.000	1.000	1.000

across all settings. Increasing k raises a marginal memory cost. For example, scaling k from 10 to 50 increases memory by just 16 MB, indicating strong memory efficiency and scalability.

- **Time Cost.** The time cost includes (1) vector-based retrieval of top- k CTI candidates, and (2) fine-grained multimodal matching for threat hunting. As illustrated in Figure 5b, latency grows linearly with k , but remains manageable. At $k = 50$, a single threat query completes in ~ 1.7 seconds on GPU and ~ 6 seconds on CPU. Reducing k to 10 lowers the GPU time to 0.4 seconds with negligible performance degradation. Overall, APT-CGLP supports real-time operation at scale. For instance, an enterprise-scale alert stream ($\sim 2,000$ events [26]) can be processed in under 15 minutes, underscoring the framework’s practicality.

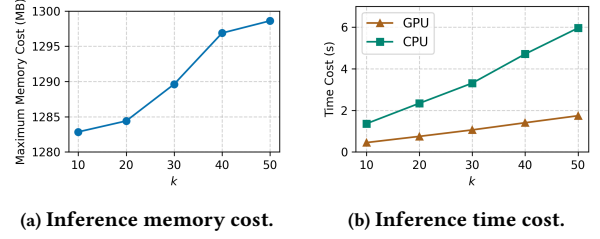
5.5 RQ4. Alert Validation Analysis

APT-CGLP can also serve as an alert validator to reduce false positives generated by upstream APT detection systems. To evaluate this capability, we leverage alerts produced by the anomaly-based detection system MAGIC [28], including both true and false alarms. Moreover, two metrics are introduced to quantify alert verification performance: 1) **Alert Filtering Rate (AFR)**—the proportion of total alerts that are successfully filtered; and 2) **Threat Retention Rate (TRR)**—the fraction of true threat alerts correctly preserved.

As reported in Table 5, APT-CGLP achieves AFRs over 90% across all four datasets, while maintaining consistently high TRRs. These results demonstrate its strong capability to eliminate spurious alerts without compromising coverage of genuine threats. Anomaly-based detectors [13, 16, 28, 48] often exhibit elevated FPRs due to rigid thresholds and reliance on behavior baselines, exacerbating the problem of alert fatigue [26] that delays or impairs incident response. As benign behavior evolves over time, these systems often misclassify normal deviations as threats [22]. We mitigate this issue by grounding alert filtering using external CTI knowledge.

6 Discussion

LLM Generalization: We tested various LLMs (e.g., Qwen2.5) and found that the proposed framework is not tied to a specific LLM.

**Figure 5: Overhead of APT-CGLP for different k .**

This is because the tasks performed in our framework (e.g., behavior extraction) are general, non-specialized tasks [9, 23] that are well within the capabilities of most LLMs, while larger models typically perform better.

Hallucination Risk. The hallucinations risk of LLMs usage in our framework is partially limited by the closed, understanding-oriented nature of our tasks and the restricted graph scales during data synthesis. Future work will investigate this risk and potential mitigation strategies, such as consistency checking [11].

Limitations. APT hunting evaluation requires datasets with paired audit data and attack reports, which, to our knowledge, only DARPA datasets can satisfy. Other datasets that either lack complete audit records (e.g., StreamSpot [35]) or do not provide detailed attack reports (e.g., LANL [45]) are not suitable for this evaluation. Although our evaluation covers more than ten distinct attack types across different platforms, the scope of benchmarks remains limited. We leave the construction of more comprehensive and diverse APT datasets, encompassing broader attack behaviors and operating environments, as an important direction for future work.

7 Conclusion

We propose APT-CGLP, an end-to-end APT hunting system that enables cross-modal semantic matching between provenance graphs and CTI reports. To mitigate data scarcity, we introduce a Graph2CTI module that employs LLM-driven heuristics to generate informative training pairs. To enhance the usability of real-world CTI reports, the CoT reasoning capacity of LLMs is employed to extract actionable insights. Moreover, we design a multi-objective training framework that integrates contrastive learning and masked modeling to achieve multi-scale cross-modal semantic alignment. Extensive experiments demonstrate that APT-CGLP delivers state-of-the-art threat hunting performance with minimal manual effort, low overhead, and high scalability.

Acknowledgements

This work has been partially supported by the National Natural Science Foundation of China (Grant Nos. U22B2028, 62372410), the Key Research Program of Hangzhou (Grant No. 2025SZZD1A56), the Zhejiang Provincial Natural Science Foundation of China (Grant No. LD22F020002), the Key Research Program of Shaoxing (Grant No. 2025B11004), the Zhejiang Provincial Natural Science Foundation of China (Grant No. LZ23F020011), and the Zhejiang Province Leading Goose Program (Grant No. 2025C01013).

References

- [1] 2017. Transparent Computing Engagement 3 Data Release. <https://github.com/darpa-i2o/Transparent-Computing/blob/master/README-E3.md>.
- [2] 2020. Operationally Transparent Cyber (OpTC) Data Release. <https://github.com/FiveDirections/OpTC-data>.
- [3] 2024. BERT base model (uncased). <https://huggingface.co/google-bert/bert-base-uncased>.
- [4] 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. <https://ai.meta.com/blog/meta-llama-3/>.
- [5] Enes Altinisik, Fatih Deniz, and Hüsrev Taha Sencar. 2023. ProvG-Searcher: A Graph Representation Learning Approach for Efficient Provenance Graph Search. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*. 2247–2261. doi:10.1145/3576915.3623187
- [6] Ahmed Aly, Shahrear Iqbal, Amr Youssef, and Essam Mansour. 2024. MEGRAPT: A Memory-Efficient APT Hunting System Based on Attack Representation Learning. *IEEE Transactions on Information Forensics and Security* (2024).
- [7] Md Monowar Anjum, Shahrear Iqbal, and Benoit Hamelin. 2021. Analyzing the usefulness of the DARPA OpTC dataset in cyber threat detection research. In *Proceedings of the 26th ACM Symposium on Access Control Models and Technologies*. 27–32.
- [8] Daniel Arp, Erwin Quiring, Feargus Pendlebury, Alexander Warnecke, Fabio Pierazzi, Christian Wressneger, Lorenzo Cavallaro, and Konrad Rieck. 2022. Dos and don'ts of machine learning in computer security. In *31st USENIX Security Symposium (USENIX Security 22)*. 3971–3988.
- [9] Dipkamal Bhushal, Md Tanvirul Alam, Le Nguyen, Ashim Mahara, Zachary Lightcap, Rodney Frazier, Romy Fieblinger, Grace Long Torales, Benjamin A Blakely, and Nidhi Rastogi. 2024. SECURE: Benchmarking Large Language Models for Cybersecurity. In *2024 Annual Computer Security Applications Conference (ACSAC)*. IEEE, 15–30.
- [10] Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).
- [11] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. In *The Twelfth International Conference on Learning Representations*.
- [12] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, et al. 2025. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*. Springer, 370–387.
- [13] Tieming Chen, Chengyu Dong, Mingqi Lv, Qijie Song, Haiwen Liu, Tiantian Zhu, Kang Xu, Ling Chen, Shouling Ji, and Yuan Fan. 2022. APT-KGL: An Intelligent APT Detection System Based on Threat Knowledge and Heterogeneous Provenance Graph Learning. *IEEE Transactions on Dependable and Secure Computing* (2022), 1–15. doi:10.1109/TDSC.2022.3229472
- [14] Wenrui Cheng, Tiantian Zhu, Tieming Chen, Qixuan Yuan, Jie Ying, Hongmei Li, Chunlin Xiong, Mingda Li, Mingqi Lv, and Yan Chen. 2024. CRUcialG: Reconstruct Integrated Attack Scenario Graphs by Cyber Threat Intelligence Reports. *arXiv preprint arXiv:2410.11209* (2024).
- [15] Zijun Cheng, Rujie Dai, Lei Qi Wang, Ziyang Yu, Qiujian Lv, Yan Wang, and Degang Sun. 2023. GHunter: A Fast Subgraph Matching Method for Threat Hunting. In *2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. 1014–1019. doi:10.1109/CSCWD57460.2023.10152818
- [16] Zijun Cheng, Qiujian Lv, Jinyuan Liang, Yan Wang, Degang Sun, Thomas Pasquier, and Xueyuan Han. 2023. Kairos: Practical Intrusion Detection and Investigation using Whole-system Provenance. *arXiv preprint arXiv:2308.05034* (2023).
- [17] MITRE Corporation. 2015. MITRE ATT&CK. <https://attack.mitre.org>.
- [18] DavidJBianco. 2014. The Pyramid of Pain. <https://detect-respond.blogspot.com/2013/03/the-pyramid-of-pain.html>.
- [19] Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [20] Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. 2022. Translation between Molecules and Natural Language. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 375–413.
- [21] Carl Edwards, ChengXiang Zhai, and Heng Ji. 2021. Text2mol: Cross-modal molecule retrieval with natural language queries. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 595–607.
- [22] Pengcheng Fang, Peng Gao, Changlin Liu, Erman Ayday, Kangkook Jee, Ting Wang, Yanfang Fanny Ye, Zhuotao Liu, and Xusheng Xiao. 2022. {Back-Propagating} system dependency impact for attack investigation. In *31st USENIX Security Symposium (USENIX Security 22)*. 2461–2478.
- [23] Romy Fieblinger, Md Tanvirul Alam, and Nidhi Rastogi. 2024. Actionable cyber threat intelligence using knowledge graphs and large language models. In *2024 IEEE European symposium on security and privacy workshops (EuroS&PW)*. IEEE, 100–111.
- [24] Peng Gao, Fei Shao, Xiaoyuan Liu, Xusheng Xiao, Zheng Qin, Fengyuan Xu, Prateek Mittal, Sanjeev R Kulkarni, and Dawn Song. 2021. Enabling efficient cyber threat hunting with cyber threat intelligence. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. IEEE, 193–204.
- [25] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints* (2023).
- [26] Wajih Ul Hassan, Shengjian Guo, Ding Li, Zhengzhang Chen, Kangkook Jee, Zhichun Li, and Adam Bates. 2019. NoDoze: Combatting Threat Alert Fatigue with Automated Provenance Triage. *Network and Distributed Systems Security Symposium* (Feb. 2019).
- [27] Yupeng Hou, Binbin Hu, Wayne Xin Zhao, Zhiqiang Zhang, Jun Zhou, and Ji-Rong Wen. 2022. Neural graph matching for pre-training graph neural networks. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*. SIAM, 172–180.
- [28] Zian Jia, Yun Xiong, Yuhong Nan, Yao Zhang, Jinjing Zhao, and Mi Wen. 2023. MAGIC: Detecting Advanced Persistent Threats via Masked Graph Representation Learning. (2023).
- [29] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.
- [30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.
- [31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.
- [32] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.
- [33] Zhenyuan Li, Jun Zeng, Yan Chen, and Zhenkai Liang. 2022. AttackKG: Constructing technique knowledge graph from cyber threat intelligence reports. In *European Symposium on Research in Computer Security*. Springer, 589–609.
- [34] Mingqi Lv, HongZhe Gao, Xuebo Qiu, Tieming Chen, Tiantian Zhu, Jinyin Chen, and Shouling Ji. 2024. TREC: APT Tactic/Technique Recognition via Few-Shot Provenance Subgraph Learning. *arXiv preprint arXiv:2402.15147* (2024).
- [35] Emaad Manzoor, Sadeq M Milajerdi, and Leman Akoglu. 2016. Fast memory-efficient anomaly detection in streaming heterogeneous graphs. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1035–1044.
- [36] microsoft. [n. d.]. <https://www.microsoft.com/security/blog/topic/threat-intelligence/>.
- [37] Sadeq M. Milajerdi, Birhanu Eshete, Rigel Gjomemo, and V.N. Venkatakrishnan. 2019. POIROT: Aligning Attack Behavior with Kernel Audit Records for Cyber Threat Hunting. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 1795–1812. doi:10.1145/3319535.3363217
- [38] Eric Peterson. 2024. Advanced Persistent Threats: A Growing Danger to Enterprise Security. <https://www.linkedin.com/pulse/advanced-persistent-threats-growing-danger-enterprise-dyo4e>.
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [40] Mati Ur Rehman, Hadi Ahmadi, and Wajih Ul Hassan. 2024. FLASH: A Comprehensive Approach to Intrusion Detection via Provenance Graph Representation Learning. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 139–139.
- [41] SANS and Qualys. 2019. SANS 2018 Threat Hunting Survey Results. <https://www.qualys.com/forms/whitepapers/sans-2018-threat-hunting-survey-results/>.
- [42] Kiavash Satvat, Rigel Gjomemo, and VN Venkatakrishnan. 2021. Extractor: Extracting attack behavior from threat reports. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 598–615.
- [43] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning With Neural Tensor Networks for Knowledge Base Completion. In *Advances in Neural Information Processing Systems*, C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (Eds.), Vol. 26. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2013/file/b337e84de8752b27eda3a12363109e80-Paper.pdf
- [44] Bing Su, Dazhao Du, Zhao Yang, Yujie Zhou, Jiangmeng Li, Anyi Rao, Hao Sun, Zhiwu Lu, and Ji-Rong Wen. 2022. A molecular multimodal foundation model associating molecule graphs with natural language. *arXiv preprint arXiv:2209.05481* (2022).
- [45] Melissa JM Turcotte, Alexander D Kent, and Curtis Hash. 2019. Unified host and network data set. In *Data science for cyber-security*. World Scientific, 1–22.
- [46] virustotal. [n. d.]. <https://blog.virustotal.com/>.
- [47] Thomas D Wagner, Khaled Mahbub, Esther Palomar, and Ali E Abdallah. 2019. Cyber threat intelligence sharing: Survey and research directions. *Computers & Security* 87 (2019), 101589.
- [48] Su Wang, Zhiliang Wang, Tao Zhou, Hongbin Sun, Xia Yin, Dongqi Han, Han Zhang, Xingang Shi, and Jiahai Yang. 2022. THREATTRACE: Detecting and Tracing Host-Based Threats in Node Level Through Provenance Graph Learning. *IEEE*

- Transactions on Information Forensics and Security* 17 (2022), 3972–3987. doi:10.1109/TIFS.2022.3208815
- [49] Weihang Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. 2024. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems* 37 (2024), 121475–121499.
- [50] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. Heterogeneous graph attention network. In *The world wide web conference*. 2022–2032.
- [51] Renzheng Wei, Lijun Cai, Lixin Zhao, Aimin Yu, and Dan Meng. 2021. DeepHunter: A Graph Neural Network Based Approach for Robust Cyber Threat Hunting. In *Security and Privacy in Communication Networks*. 3–24. doi:10.1007/978-3-030-90019-9_1
- [52] Zhihao Wen and Yuan Fang. 2023. Augmenting low-resource text classification with graph-grounded pre-training and prompting. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 506–516.
- [53] Chunlin Xiong, Tiantian Zhu, Weihao Dong, Linqi Ruan, Runqing Yang, Yueqiang Cheng, Yan Chen, Shuai Cheng, and Xutong Chen. 2022. Conan: A Practical Real-Time APT Detection System With High Accuracy and Efficiency. *IEEE Transactions on Dependable and Secure Computing* 19, 1 (Jan. 2022), 551–565. doi:10.1109/TDSC.2020.2971484
- [54] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826* (2018).
- [55] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917* (2022).
- [56] Mengmei Zhang, Mingwei Sun, Peng Wang, Shen Fan, Yanhu Mo, Xiaoxiao Xu, Hong Liu, Cheng Yang, and Chuan Shi. 2024. GraphTranslator: Aligning Graph Model to Large Language Model for Open-ended Tasks. In *Proceedings of the ACM on Web Conference 2024*. 1003–1014.
- [57] Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B Hashimoto. 2024. Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics* 12 (2024), 39–57.
- [58] Yongheng Zhang, Tingwen Du, Yunshan Ma, Xiang Wang, Yi Xie, Guozheng Yang, Yuliang Lu, and Ee-Chien Chang. 2024. AttackKG+: Boosting Attack Knowledge Graph Construction with Large Language Models. *arXiv preprint arXiv:2405.04753* (2024).
- [59] Yimei Zhang, Guojian Shen, Kaili Ning, Tongwei Ren, Xuebo Qiu, Mengmeng Wang, and Xiangjie Kong. 2025. Improving Region Representation Learning from Urban Imagery with Noisy Long-Caption Supervision. *arXiv preprint arXiv:2511.07062* (2025).
- [60] Zheng Zhang, Yuntong Hu, Bo Pan, Chen Ling, and Liang Zhao. 2024. TAGA: Text-Attributed Graph Self-Supervised Learning by Synergizing Graph and Text Mutual Transformations. *arXiv preprint arXiv:2405.16800* (2024).
- [61] Tiantian Zhu, Jinkai Yu, Chunlin Xiong, Wenrui Cheng, Qixuan Yuan, Jie Ying, Tieming Chen, Jiabo Zhang, Mingqi Lv, Yan Chen, Ting Wang, and Yuan Fan. 2023. APTSHIELD: A Stable, Efficient and Real-Time APT Detection System for Linux Hosts. *IEEE Transactions on Dependable and Secure Computing* (2023), 1–18. doi:10.1109/TDSC.2023.3243667
- [62] Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. 2024. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024).

Appendix

A Activity Subgraph Sampling

We design a subgraph sampling algorithm to extract semantic-rich subgraphs from provenance graphs, serving as templates for generating paired CTI reports to build a cross-modal learning dataset. As shown in Algorithm 1, it adopts a three-layer Breadth-First Search (BFS) strategy and takes as input a provenance graph pg and a size range $[minN, maxN]$, and outputs a set of subgraphs SG_s .

To better approximate real attack scenarios, we design a progressive sampling process. In Layer 1 (lines 6-7), the algorithm samples process-type neighbors from socket nodes to capture initial attack entry points. In Layer 2 (lines 8-10), it performs balanced sampling across different node types through BFS expansion, ensuring structural diversity in the subgraph. In Layer 3 (lines 11-14), the subgraph is conditionally filled when its size falls below the

minimum threshold, with additional nodes randomly sampled to reach a target size within the specified range. Finally, each sampled subgraph is validated to ensure it contains all node types and falls within the predefined size range (lines 15-16), guaranteeing representativeness of genuine attack patterns.

Algorithm 1: Activity Subgraph Sampling Algorithm

Input: Provenance graph pg ; Size range $[minN, maxN]$

Output: Subgraphs SG_s

```

1  $SG_s \leftarrow []$ 
2  $nei\_map \leftarrow PrecomputeNeighbors(pg)$ 
3  $seed\_nodes \leftarrow FilterByType(pg, "socket")$ 
4 foreach  $v \in seed\_nodes$  do
5    $sg \leftarrow \{v\}$ 
6   // Layer 1: Sample process neighbors
7    $N_1 \leftarrow SampleByType(nei\_map[v], "process")$ 
8    $sg.update(N_1)$ 
9   // Layer 2: Balanced sampling across node types
10   $N_2 \leftarrow BFS(N_1, nei\_map, sg)$ 
11   $N_2^{sampled} \leftarrow TypeBalancedSample(N_2)$ 
12   $sg.update(N_2^{sampled})$ 
13  // Layer 3: Conditionally fill subgraph
14  if  $|sg| < minN$  then
15     $N_3 \leftarrow BFS(N_2^{sampled}, nei\_map, sg)$ 
16     $N_3^{sampled} \leftarrow RandomSample(N_3, minN - |sg|, maxN - |sg|)$ 
17     $sg.update(N_3^{sampled})$ 
18  // Validate subgraph
19  if  $minN \leq |sg| \leq maxN$  &  $CheckNodeTypes(sg)$  then
20     $SG_s.add(sg)$ 
21 end
22 return  $SG_s$ 

```

B CTI Generation Prompt

We harness LLaMA’s in-context learning capabilities to construct synthetic training data, thereby enriching cross-modal training datasets. This approach facilitates the generation of desired CTI-like reports by offering specific demonstrations within the prompt. We customize 5 various in-context learning demonstrations to guide the LLM in generating sufficient diverse synthetic reports, enabling the model to acquire robustness and generalization capabilities during training. A prompt example is illustrated in Figure 6.

C CTI Denoising Prompt

We employ LLaMA with Chain-of-Thought prompting to extract attack-relevant information from unstructured CTI reports. This method reduces noise and produces coherent, actionable narratives suitable for downstream threat analysis. A prompt example illustrated in Figure 7 comprises three main parts: task description, multi-step processing instructions, and input data.

D Graph Isomorphism Network

We employ the Graph Isomorphism Network (GIN) [54] as graph encoder for its strong expressive power. The encoding process

Task:

You are a professional cyber threat analyst. Your task is to convert structured behavioral interactions from audit logs into a clear and natural language narrative. Please refer to the following examples for guidance:

Input:

1. Process D executed file F.
2. Process D executed file G.
3. Process D read from file A.
4. Process D connected to socket L and exchanged data.
5. Process C modified attributes of file J and created process K.

Output:

Here is the converted output: Process D initiated by executing files F and G. It then read data from file A and established a connection with socket L. Meanwhile, Process C modified the attributes of file J and created process K.

Note: Please convert the following structured behavior into a natural language description except repetitive, immediate actions and the initial self-execution/creation of the process itself. Output the converted result directly, without any formats.

Input: {structured_interactions}

Figure 6: In-Context learning prompt for report generation.

Task:

You are a cybersecurity analyst specializing in behavioral sequence mapping. Your task is to transform complex threat reports into a highly condensed, action-oriented behavioral summary. You must process the threat report by explicitly showing your analysis in three distinct steps.

Step 1. Entity Identification:

Isolate three primary entity types: Processes (executables, such as "*.exe"), Files (only filenames without directory path, such as "*.dll"), and Sockets (IPs, such as "127.0.0.1"). Ignore all other named entities (e.g., organizations, threat actors, API names, general terms).

Step 2. Interaction Extraction:

For each process in Step 1, extract its actions with other objects identified in Step 1. Map every action to: [read, write, connect, send, receive, create, delete, modify, clone, execute], and must form a clear behavioral triplet (Subject-Action-Object).

Step 3. Knowledge Distillation:

Integrate the triplets in Step 2 into a final summary using the mandatory format: "Summary Sentence: Process [Name] [Action] [Object]. [Action] [Object], and [Action] [Object]. Process [Name]...".

You must display the results of Step 1, Step 2, and Step 3 clearly. You only need to output the content I specify for you, without introductory text, conclusion, and "Note" sections.

Input Report: {CTI_report}

Figure 7: Chain-of-Thought prompt for CTI denoising.

comprises aggregation, combination, and update phases, which iteratively refine node features by integrating semantic signals from their neighbors. We initialize node features $\{h_v^{(0)} | v \in \mathcal{V}\}$ using a text encoder applied to node types and names, while edge features $\{e_{uv} | u \in \mathcal{N}(v)\}$ are derived from their interaction types. At layer k , each node v aggregates information from its neighbors through a learnable message function $g(\cdot)$:

$$m_v^{(k)} = \sum_{u \in \mathcal{N}(v)} g(h_u^{(k-1)}, e_{uv}).$$

A trainable coefficient $\epsilon^{(k)}$ then controls the relative contribution of the node's previous representation and the aggregated neighborhood message:

$$s_v^{(k)} = (1 + \epsilon^{(k)}) \cdot h_v^{(k-1)} + \text{agg}_v^{(k)}.$$

The updated node representation is produced by applying a layer-specific MLP:

$$h_v^{(k)} = \text{MLP}^{(k)}(s_v^{(k)}).$$

After k iterations, each node embedding encodes information from its k -hop neighborhood, effectively capturing local structural and semantic context.