DiffuPac: Contextual Mimicry in Adversarial Packets Generation via Diffusion Model

Anonymous Author(s)
Affiliation
Address
email

Abstract

In domains of cybersecurity, recent advancements in Machine Learning (ML) and Deep Learning (DL) have significantly enhanced Network Intrusion Detection Systems (NIDS), improving the effectiveness of cybersecurity operations. However, attackers have also leveraged ML/DL to develop sophisticated models that generate adversarial packets capable of evading NIDS detection. Consequently, defenders must study and analyze these models to prepare for the evasion attacks that exploit NIDS detection mechanisms. Unfortunately, conventional generation models often rely on unrealistic assumptions about attackers' knowledge of NIDS components, making them impractical for real-world scenarios. To address this issue, we present DiffuPac, a first-of-its-kind generation model designed to generate adversarial packets that evade detection without relying on specific NIDS components. DiffuPac integrates a pre-trained Bidirectional Encoder Representations from Transformers (BERT) with diffusion model, which, through its capability for conditional denoising and classifier-free guidance, effectively addresses the real-world constraint of limited attacker knowledge. By concatenating malicious packets with contextually relevant normal packets and applying targeted noising only to the malicious packets, DiffuPac seamlessly blends adversarial packets into genuine network traffic. Through evaluations on real-world datasets, we demonstrate that DiffuPac achieves strong evasion capabilities against sophisticated NIDS, outperforming conventional methods by 8.83 percentage points while preserving the attack functionality and practicality of the generated adversarial packets.

1 Introduction

Network Intrusion Detection Systems (NIDS) play a pivotal role in safeguarding the vast array of digital devices and infrastructures that permeate our lives. As of 2023, the global count of active IoT devices is expected to reach approximately 15.14 billion, with projections suggesting a rise to 30 billion by 2030 (Statista [2023]). This explosive growth, fueled by applications spanning from consumer electronics to industrial automation and healthcare, presents formidable security challenges. To meet these challenges, advancements in Machine Learning (ML) and Deep Learning (DL) have significantly bolstered the efficacy of NIDS in monitoring IoT traffic and detecting malicious activities (Talaei Khoei and Kaabouch [2023]; Talaei Khoei et al. [2023]).

However, the rapid evolution of generative AI technologies has ushered in a new era of cybersecurity threats, notably through the creation of adversarial packets designed to evade detection by even the most sophisticated NIDS. These AI-driven attacks can emulate and synthesize legitimate network behaviors, presenting an unprecedented challenge to existing security paradigms. Generative models, particularly those trained on extensive datasets of genuine network traffic, can generate adversarial packets that blend malicious functionalities within seemingly normal packet sequences, thus effectively camouflaging their malicious intents.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

In response to these evolving threats, it is critical for cybersecurity defenders to deepen their understanding of these generative AI models. By scrutinizing the mechanisms through which these models generate adversarial packets, defenders can better anticipate and counteract adversarial tactics that compromise the detection capabilities of NIDS (Ibitoye et al. [2019]; Khazane et al. [2024]). The urgency to develop innovative solutions that can adapt to and preempt these adversarial tactics is paramount, ensuring the reliability and robustness of NIDS in an increasingly complex threat landscape.

Traditional methods for generating adversarial packets have primarily relied on direct engagement with NIDS or the use of surrogate classifiers, often assuming unrealistic levels of attacker access to NIDS configurations. Table 1 summarizes recent literature on adversarial packet generation, highlighting the limitations of these approaches. Studies using techniques such as Network Emulator (NetEM) and Metasploit (Homoliak et al. [2018]), Generative Adversarial Network (GAN) and Particle Swarm Optimization (PSO) (Han et al. [2021]), and Reinforcement Learning (RL) (Hore et al. [2023]) demonstrate that while these methods successfully modified the packets behavior, their efficacy in evading detection in real-world conditions remains suboptimal. The reliance on detailed knowledge of NIDS models is flawed, as attackers typically operate with limited information about the underlying security infrastructure. This gap underscores the necessity for more practical and effective adversarial generation methods that align with the realistic constraints faced by attackers.

Table 1: Summary of recent literature on adversarial packets generation.

Author	**Year**	**Data Set**	**Classifier-Techniques**	**Algorithm**
Homoliak et al. [2018]	2018	ASNM-NBPO	Surrogate Classifier	NetEM, Metasploit
Hashemi et al. [2019]	2019	CICIDS-2018	Surrogate Classifier	Trial and error
Kuppa et al. [2019]	2019	CICIDS-2018	Surrogate Classifier	Manifold Approx.
Han et al. [2021]	2021	Kitsune, CICIDS-2017	NIDS feature extractor	GAN and PSO
Sharon et al. [2021]	2021	Kitsune, CICIDS-2017	NIDS classifier	LSTM
Hore et al. [2023]	2023	CICIDS-2017, CICIDS-2018	Surrogate Classifier	RL

To address these challenges, we introduce *DiffuPac: Contextual Mimicry in Adversarial Packet Generation via Diffusion Model*, a novel solution that leverages the combined strengths of Bidirectional Encoder Representations from Transformers (BERT) and diffusion model. This innovative fusion not only promises high accuracy in generating adversarial packets but also operates under the realistic assumption that attackers lack direct access to NIDS models. DiffuPac leverages the extensive contextual understanding provided by BERT, which has been trained on diverse datasets representing a wide range of network behaviors, along with the generative capabilities of diffusion models. This fusion results in a sophisticated adversarial tactics where the elements of the attack are seamlessly integrated into the network traffic, making them indistinguishable from legitimate data. This capability represents a significant leap forward, offering a stealthy approach that outmaneuvers current NIDS through advanced mimicry rather than direct confrontation.

In summary, the principal contributions of this paper are as follows: (a) we have pioneered the integration of BERT and diffusion models to create DiffuPac, marking a first in the cybersecurity domain. This novel methodology sets a precedent in the field by blending the advanced contextual comprehension of network traffic with sophisticated generative capabilities to produce adversarial packets that are both stealthy and indistinguishable from genuine traffic; (b) we introduce a unique concatenation strategy coupled with targeted noising techniques. These innovations ensure that the adversarial packets not only blend seamlessly into the network environment but also dynamically adapt to evade modern detection systems; (c) DiffuPac advances a classifier-free approach to adversarial packet generation. This approach challenges traditional dependency on surrogate classifiers, offering a new paradigm that more accurately reflects the constraints and capabilities of real-world attackers; (d) Lastly, our extensive experimental evaluations, conducted on real-world datasets, demonstrate that DiffuPac significantly outperforms existing methods in terms of evasion effectiveness, establishing new benchmarks for the generation of adversarial packets. These contributions collectively push the boundaries of what is possible in the realm of network security, paving the way for more resilient cybersecurity defenses and a deeper understanding of adversarial tactics in network environments.

2 Related Works

Adversarial Attacks on ML/DL-based NIDS. NIDS are crucial for protecting digital infrastructures by monitoring network traffic and identifying potential threats through signature-based and anomaly-based detection paradigms. Signature-based NIDS use pattern matching against predefined threat databases, while anomaly-based NIDS employ machine learning models to detect deviations from benign traffic patterns, providing an advantage in detecting sophisticated threats. The transition to DL in anomaly-based NIDS, as highlighted by Ahmad et al. [2021], has enhanced threat detection due to the capability to learn abstract patterns, though this shift has introduced a vulnerability to adversarial attacks. These attacks modify network data subtly to evade NIDS, a phenomenon first noted in computer vision (Szegedy et al. [2014]) and now challenging DL-based NIDS. Adversarial attack tactics vary based on attacker knowledge, ranging from white-box attacks with full system knowledge to black-box attacks with no classifier knowledge, as outlined by McCarthy et al. [2022]. Black-box scenarios are common in real-world threats since attackers generally lack direct NIDS access, necessitating sophisticated modifications of data to exploit DL model vulnerabilities and blur the distinction between normal and malicious traffic.

Adversarial Attacks Evading NIDS. In the study of NIDS, researchers explore feature-level and packet-level attacks to enhance robustness. Feature-level attacks modify input network features using methods like GANs to mislead classifiers without direct knowledge of their mechanisms. For instance, Yang et al. [2018] used transfer-based and score-based attacks to deceive a Deep Neural Network (DNN) model on the NSL-KDD dataset, while Sheatsley et al. [2022] developed an Augmented JSMA (AJSMA) to ensure realistic feature modifications. However, these still fail to generate executable malicious packets since they do not provide a method for converting modified features into packet sequences. Conversely, packet-level attacks, which modify network packets directly to maintain their malicious intent while evading detection, are particularly effective. As detailed in Table 1, studies such as Hashemi et al. [2019] show successful modifications using non-payload based and mimicking operations. In the survey He et al. [2023], authors emphasized that the practicality and replayability of packet-level attacks make them more dominant, ensuring adversarial packets evade detection while remaining executable. Given this efficacy, our research model, DiffuPac, is designed to excel in this domain, outperforming previous models in robustness and detection evasion, and setting a new standard for NIDS efficacy in adversarial cybersecurity.

Diffusion Models. Represent a breakthrough in generative modeling, offering a novel framework for understanding data as a dynamic and stochastic process. Based on the works of Sohl-Dickstein et al. [2015] and Ho et al. [2020], these models conceptualize data points $\mathbf{z}_0 \in \mathbb{R}^d$ (where d is a positive integer) as the end result of a reverse Markov chain. In the forward diffusion process, data points $\mathbf{z}_0$ are gradually transformed into a noise distribution . This transformation is modeled by a Markov chain starting from the data distribution $q(\mathbf{z}_0)$ and ending in a Gaussian distribution $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where T represents the total number of timesteps in the forward diffusion process. The transitions are defined $q(\mathbf{z}_t|\mathbf{z}_{t-1}) = \mathcal{N}(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I})$ for $1 \leq t \leq T$, where β_t as the variance scale. In the reverse process, the model f_θ (usually a U-Net or a Transformer) learns to reconstruct the original data $\mathbf{z}_0$ from the noised data $\mathbf{z}_T$. It does so by iteratively estimating the parameters $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \sigma_t^2\mathbf{I})$, where μ_θ and σ_t^2 predict the mean and variance of the distribution.

BERT and Diffusion Models Approaches in Cybersecurity and Other Fields. The application of pre-trained models, such as ET-BERT (Lin et al. [2022]) and NetGPT (Meng et al. [2023]), has revolutionized understanding of network traffic by capturing the intricacies of language and interpreting complex network patterns. Meanwhile, diffusion models have emerged as powerful generative tools capable of producing high-fidelity data for applications like image generation and notably, synthetic network traffic generation. These models leverage a unique denoising training objective to closely mimic the original data distribution, making them promising for generating realistic network traffic for testing and analysis. Despite these advancements, combining the contextual understanding of pre-trained models with the generative prowess of diffusion models remains underexplored in adversarial packet generation. This integration offers a synergistic approach, merging the nuanced comprehension of network behaviors by pre-trained models with the high-quality data generation of diffusion models. By leveraging these technologies together, cybersecurity researchers can enhance the realism of simulated network environments and develop more advanced evasion tactics.

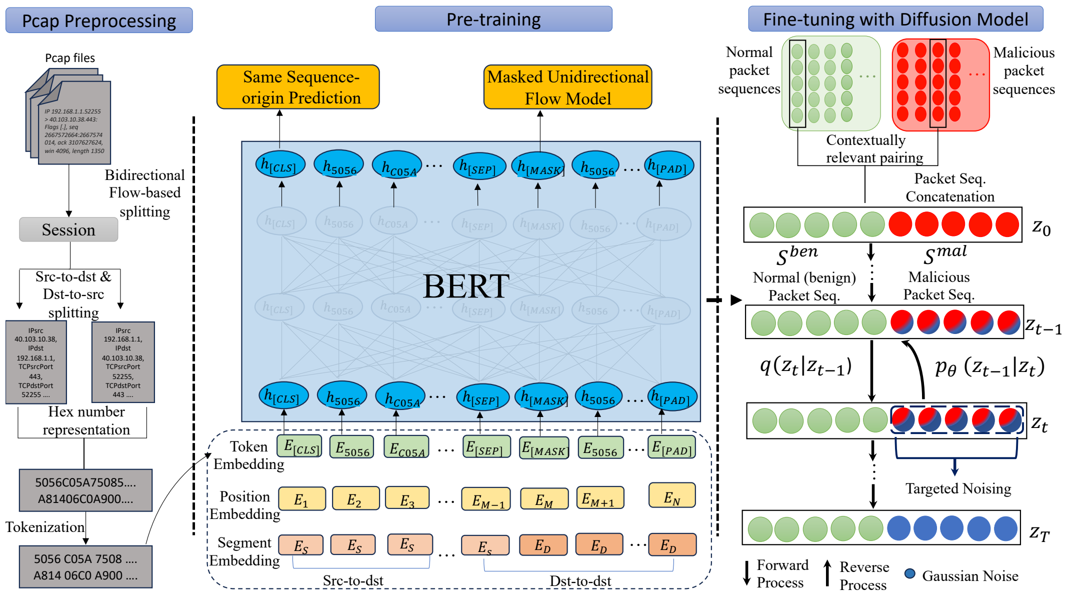


Figure 1: Proposed arhitecture is divided into three phases: pcap pre-processing, pre-training and fine-tuning with diffusion models.

3 DiffuPac

In this section, we introduce DiffuPac, a first-of-its kind adversarial packet generation model. Figure 1 illustrates the overall framework of the proposed model, consisting of three main phases: pcap pre-processing, pre-training, fine-tuning with diffusion model.

3.1 Data pre-processing

In real network environments, traffic contains diverse flows from various applications, protocols, and services, complicating the learning of stable representations. Therefore, we first split pcap (packet capture) files into sessions (bidirectional flows) based on IP addresses, port addresses, and protocols. To refine the training of the BERT model, we further split sessions into unidirectional flows, categorizing them as either source-to-destination (src-to-dst) packet sequence or destination-to-source (dst-to-src) packet sequence. This categorization is crucial for the BERT model's pre-training. Network traffic varies due to diverse protocols and network services, resulting in different formats and patterns. To handle these variations and encoding requirements, we convert each byte to its corresponding hex number and tokenize using WordPiece (Wu et al. [2016]). Each token ranges from 0 to 65535, with a dictionary size of 65536. We also incorporate special tokens `[CLS]`, `[SEP]`, `[PAD]`, and `[MASK]` for training tasks. `[CLS]` is used at the beginning of each sequence and helps in classification tasks. `[SEP]` separates different sequences or segments within the same input. `[PAD]` ensures sequences are of uniform length and satisfy the minimum length requirement. `[MASK]` is used in pre-training task, where it temporarily substitutes tokens to be predicted.

3.2 Pre-training

As shown in Figure 1, the input tokens are processed using an embedding strategy that involves the sum of three types of embeddings: token embeddings, positional embeddings, and segment embeddings. Each embedding has a dimension of 768. *Token Embeddings* are high-dimensional vectors that uniquely represent each token, acting as their exclusive identifiers. *Positional Embeddings* are used to capture the temporal relationships of tokens, ensuring that the model learns to focus on the order of data transmission. *Segment Embeddings* differentiate packets within a single flow, as packets may not inherently share semantic associations, and preserving the order of packets to maintain the temporal sequence of events in a session flow.

We design our pre-training tasks based on the approach described in Devlin et al. [2019]. Our two proposed pre-training tasks aim to capture the contextual relationships between traffic bytes. The first task involves predicting masked tokens to learn the underlying patterns and dependencies within the traffic data. The second task predicts the transmission order by determining whether packets belong to src-to-dst or dst-to-src sequences, thereby capturing the directional flow of the network traffic.

Masked Unidirectional Flow Model. Inspired by BERT's Masked Language Model (MLM) in natural language processing, we adapt this approach for network traffic analysis through our Masked Unidirectional Flow Model. This model is designed to understand and predict the semantic patterns within bidirectional network flows; that are src-to-dst sequence and dst-to-src sequence. During pre-training, each token in the input sequence is masked with a probability of 15%. Amongst these masked tokens 80% are replaced with a `[MASK]` token, 10% are replaced with a random token from the vocabulary, and 10% are left unchanged. This introduces variability that mimics real-world data inconsistencies. The training objective is to minimize the negative log-likelihood of correctly predicting the original tokens at the masked positions. Formally, the loss function L_{MUM} for this task is defined as:

$$L_{\text{MUM}} = -\sum_{i=1}^{n} 1(m_i) \log P(v_i \mid V_{\text{masked}}; \theta), \tag{1}$$

where n is the total number of tokens in the sequence, $1(m_i)$ is an indicator function that is 1 if the token v_i was masked (0 otherwise). m_i indicates the masking status of the i-th token and V_{masked} is the masked sequence. v_i is the actual token at position i and $P(v_i \mid V_{\text{masked}}; \theta)$ is the probability of predicting the original token v_i for given masked sequence and model parameters θ. The transformer encoder, characteristic of BERT, processes V_{masked} to predict each masked token. This architecture leverages self-attention mechanisms to capture dependencies and context effectively, which is crucial for understanding complex patterns in network traffic.

Same Sequence-origin Prediction. Inspired by Lin et al. [2022], this task employs a dedicated binary classifier to determine the directional origin of network traffic, specifically whether packets in a sequence originate from src-to-dst or dst-to-src. This classification enhances the model's understanding of network flows and the contextual relationships between packets, which is crucial for recognizing communication patterns and dependencies in network traffic. For this task, each packet a_r is paired with another packet a_s. 50% of the time, a_s is the next logical packet in the flow (src-to-dst or dst-to-src); the other 50% of the time, it is a randomly chosen packet from the opposite flow. The classifier then predicts whether a_s follows a_r in the correct directional sequence. This pairing strategy improves the model's capability to discern patterns in packet flows. Let $A = \{(a_r, a_s)\}$ be the set of packet pairs, where each pair is labeled with $b_w \in \{0, 1\}$ (1 if a_s follows a_r in the correct flow, 0 otherwise). The loss function L_{SSP} for this task can be formulated as:

$$L_{\text{SSP}} = -\sum_{w=1}^{K} b_w \log P(b_w = 1 \mid a_r, a_s; \theta) + (1 - b_w) \log(1 - P(b_w = 1 \mid a_r, a_s; \theta)), \tag{2}$$

where K is the total number of packet pairs and θ represents the trainable parameters of the classifier. $P(b_w = 1 \mid a_r, a_s; \theta)$ is the probability predicted by the classifier that a_s correctly follows a_r in the given network flow direction.

In summary, the final pre-training objective is the sum of the above two losses, which can defined as:

$$L = L_{\text{SSP}} + L_{\text{MUM}}. \tag{3}$$

Fu et al. [2021] mentioned that that the number of packets within each flow can vary significantly. Given the constraints imposed by packet size and the potential volume of network traffic, computational efficiency is a paramount concern. In addressing this problems, as demonstrated in Dai et al. [2023] that the initial packets in a flow contain the most significant information, we limit our analysis to the first three packets of each heavy flow.

3.3 Fine-tuning with Diffusion Models

Forward Process with Packet Sequences Concatenation Strategy. In the fine-tuning phase, the goal is to train the model so that malicious packets can mimic normal packets to bypass NIDS. This involves using packet sequences that are only from the src-to-dst, reflecting the adversary's control over the packets being sent (Hore et al. [2023]). The forward process of our diffusion model begins by embedding both normal (benign) packet sequences $\mathbf{S}^{\text{ben}}$ and malicious packet sequences $\mathbf{S}^{\text{mal}}$. This transformation of discrete packet data into a continuous feature spaces uses an embedding function adapted from Li et al. [2022]. Building upon the groundwork of DiffuSeq (Gong et al.

[2023]) approach, which typically involves random merging, our model innovates by leveraging the deep contextual insights provided by the pre-trained BERT model (Details in A.2). This allows us to strategically pair normal and malicious sequences with strong contextual alignments that show similarity in network behavior patterns. These contextually aligned pairs are crucial for mimicking normal traffic and increasing the chances of bypassing NIDS. By integrating these contextually aligned pairs into our diffusion process, the model extends the original forward chain to a new Markov transition $q_\phi(\mathbf{z}_0 \mid \mathbf{S}^{\mathrm{ben}\oplus\mathrm{mal}}) = \mathcal{N}(\mathrm{EMB}(\mathbf{S}^{\mathrm{ben}\oplus\mathrm{mal}}), \beta_0\mathbf{I})$, where $\mathrm{EMB}(\mathbf{S}^{\mathrm{ben}\oplus\mathrm{mal}})$ symbolizes the embedding transformation and concatenation of normal and malicious packet sequences.

Targeted Noising. To enhance our diffusion model's capability for adversarial packet generation, we simplify the model's state transitions by defining $\mathbf{z}_t = \mathbf{x}_t \oplus \mathbf{y}_t$, where $\mathbf{x}_t$ and $\mathbf{y}_t$ correspond to the portions of $\mathbf{z}_t$ that belong to $\mathbf{S}^{\mathrm{ben}}$ and $\mathbf{S}^{\mathrm{mal}}$, respectively. This setup allows us to strategically inject noise into only the malicious packet sequences (represented by $\mathbf{y}_t$), rather than the entire state $\mathbf{z}_t$, during each forward step $q(\mathbf{z}_t \mid \mathbf{z}_{t-1})$. This targeted noising, inspired by DiffuSeq, is pivotal for adapting conventional diffusion models for targeted modification.

Reverse Process With Normal Packet Guidance. Building upon the foundational principles of DiffuSeq, our model introduces an innovative reverse process tailored for more nuanced handling of network traffic data. This process distinctively utilizes normal packet sequences as a guiding framework, enabling a sophisticated denoising technique that treats the concatenated sequences of normal and noise-added malicious packets as a unified unit. This approach effectively "teaches" the model to perceive these malicious elements as integral parts of the normal traffic pattern. The conditional denoising process effectively employs Bayesian inference to parameterize the transition probabilities between states, ensuring precise control over each step in the reverse process. These parameterizations are mathematically articulated through the equations: $p_\theta(\mathbf{z}_{0:T}) := p(\mathbf{z}_T)\prod_{t=1}^{T} p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t)$ and $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t) = \mathcal{N}(\mathbf{z}_{t-1}; \mu_\theta(\mathbf{z}_t, t), \sigma_\theta(\mathbf{z}_t, t)$, where $\mu_\theta(\cdot)$ and $\sigma_\theta(\cdot)$ model the mean and variance necessary for the stochastic reverse transition from $\mathbf{z}_t$ to $\mathbf{z}_{t-1}$, respectively. To rigorously ensure that the malicious packets are not only recovered but also convincingly mimic normal packets, our model optimizes a specially formulated variational lower bound L_{VLB}. This objective underscores the model's effectiveness in blending malicious packets within normal packets:

$$\min_\theta L_{VLB} = \min_\theta \left[\sum_{t=2}^{T} \|\mathbf{z}_0 - f_\theta(\mathbf{z}_t, t)\|^2 + \|\mathrm{EMB}(\mathbf{S}^{\mathrm{ben}\oplus\mathrm{mal}}) - f_\theta(\mathbf{z}_1, 1)\|^2 - \log p_\theta(\mathbf{S}^{\mathrm{ben}\oplus\mathrm{mal}} \mid \mathbf{z}_0) \right] \tag{4}$$

This objective is particularly focused on accurately reconstructing the initial state $\mathbf{z}_0$ from the noised states, with a distinct emphasis on ensuring that the malicious components are seamlessly integrated into the normal packets pattern.

Preserving Packets' Integrity. We propose an innovative approach that utilizes a parallel data structure called *"model dictionary"* that clearly distinguishes between mutable and immutable fields within each packet. By categorizing fields into mutable (e.g. TCP flags, TTL, and window size) and immutable (including the critical 5-tuple information and payload), we ensure that modifications during training do not compromise the packet's integrity. The model dictionary serves a dual purpose: it retains the original values of immutable fields and establishes a link to their mutable counterparts. This approach allows for dynamic modification of mutable fields during training without affecting the core attributes of the packet. After training, any changes made to mutable fields are seamlessly integrated with the preserved immutable fields. This recombination ensures that the modified packets maintain their operational integrity and are indistinguishable from genuine traffic in real-world scenarios. By bypassing traditional domain constraints—which typically complicate training and pose convergence challenges—our method simplifies the training process and obviates the need for additional compensatory loss functions (Letcher [2021]). The outcome is a highly effective generation of adversarial packets that are capable of evading detection without compromising the essential characteristics of the original packets.

4 Experiments

We conduct experiments to validate the performance of DiffuPac on 6 types of attacks, against 6 classifiers.

Table 2: Comparative analysis of attack detection and evasion rates.

(a) Botnet

Feature Extractor	Classifier	Detection P	Detection R	Detection F1	Evasive (*MER*) Baseline	Evasive (*MER*) **Ours**
CIC FLowMeter	KitNET	0.82	0.94	0.90	35.54%	**42.32%**
	DT	0.79	0.90	0.82	32.77%	**59.73%**
	IF	0.99	0.90	0.94	39.15%	**43.71%**
	MLP	0.92	0.82	0.86	39.72%	**61.25%**
	SVM	0.99	0.90	0.94	**88.19%**	63.19%
	LR	0.83	0.91	0.89	23.15%	**41.12%**
AfterImage	KitNET	0.96	0.90	0.94	**99.18%**	74.46%
	DT	0.79	0.90	0.84	63.42%	**72.13%**
	IF	0.99	0.90	0.94	31.48%	**52.79%**
	MLP	0.96	0.97	0.97	48.60%	**64.92%**
	SVM	0.99	0.90	0.94	40.31%	**69.19%**
	LR	0.96	0.90	0.93	53.28%	**58.98%**

(b) MITM

Feature Extractor	Classifier	Detection P	Detection R	Detection F1	Evasive (*MER*) Baseline	Evasive (*MER*) **Ours**
CIC FLowMeter	KitNET	0.91	0.94	0.90	37.31%	**53.26%**
	DT	0.74	0.79	0.76	49.77%	**61.23%**
	IF	0.99	0.90	0.94	25.95%	**47.71%**
	MLP	0.76	0.71	0.74	51.43%	**70.47%**
	SVM	0.73	0.76	0.78	41.09%	**60.19%**
	LR	0.73	0.76	0.71	36.45%	**44.72%**
AfterImage	KitNET	0.94	0.96	0.93	**68.79%**	58.48%
	DT	0.75	0.89	0.84	53.18%	**70.04%**
	IF	0.81	0.83	0.86	26.53%	**45.71%**
	MLP	0.92	0.90	0.93	50.65%	**71.45%**
	SVM	0.99	0.90	0.94	63.51%	**66.53%**
	LR	0.91	0.94	0.90	44.68%	**52.38%**

4.1 Experimental Setup

Dataset. The datasets used for this model are Kitsune Dataset (Mirsky et al. [2018]) and CICIDS-2017 Dataset (Sharafaldin et al. [2018]). Initially, *pre-training of the BERT model* utilizes a large subset of unlabeled network traffic to leverage the model's capability to capture diverse traffic patterns, accounting for 60% of the total data. *Fine-tuning phase*, the focus shifts to a smaller, labeled dataset, which constitutes 20% of the total data. This dataset is distinctly partitioned into malicious and normal packets, which is crucial for training the model to mimic malicious packets as normal. *Training the classifier and NIDS* then utilizes another 10% of the total data consisting of labeled portion. *Testing phase* is conducted with the remaining 10% of the data, reserved exclusively for evaluating the model's efficacy. This phase includes testing both original and mimicked malicious packets to rigorously assess the model's real-world applicability and its capability to generalize across unseen data.

Baseline Model. We choose Traffic Manipulator (Han et al. [2021]) as the baseline whereby the model formulates the attack as a bi-level optimization problem. The first level of optimization is to find adversarial features closest to the original feature via GAN. The second level focuses on packet modifications such as delaying, splitting, or injecting dummy packets, all tailored to exploit features recognized by the NIDS. This method, however, often assumes access to NIDS feature extractors, making it impractical in real-world scenarios where attackers typically lack such insider knowledge. Conversely, DiffuPac operates under the assumption of zero prior knowledge about NIDS configurations or feature extractors.

Implementation Details. DiffuPac leverages a BERT model with 12 transformer blocks, each featuring 12 attention heads and an embedding dimension of 768, to capture nuanced relationships in network traffic. This setup supports the maximum sequence length of 512. The diffusion model incorporates time step embedding to provide temporal context, enhancing the model's understanding of the reverse process. To manage the computational demands, particularly during the sampling phase, a projection layer reduces the embedding dimension to 128. Additionally, DiffuPac utilizes the FAISS library (Douze et al. [2024]), renowned for its capability to handle large-scale similarity search and clustering of dense vectors.

Experimental Details. Our experimental framework integrates advanced feature extraction tools and a diverse set of machine learning classifiers, building on methodologies established in the Traffic Manipulator research. We employ 2 feature extractors: AfterImage (Mirsky et al. [2018]), which provides detailed packet statistics, and CICFlowMeter (Draper-Gil et al. [2016]), which assesses connection-level metrics. These tools offer a comprehensive view of network traffic, capturing both packet-level and flow-level data essential for nuanced analysis. These feature extractors are integral to NIDS detection as they preprocess network data into structured features that are then analyzed by ML classifiers. This preprocessing step is crucial for transforming raw network traffic into a format that classifiers can effectively interpret, thereby enhancing the accuracy of the anomaly detection. We utilize a variety of machine learning classifiers for anomaly detection, including *KitNET* (Mirsky et al. [2018]), an ensemble of autoencoders; *Multi-Layer Perceptron (MLP)* for deep learning; *Logistics Regression (LR)*, *Decision Tree (DT)*, and *Support Vector Machine (SVM)* for traditional approaches; and *Isolation Forest (IF)* for outlier detection. Our experiments cover 6 types of attacks—*Man-in-the-*

Middle (MITM), *Botnet*, *Brute Force*, *DDoS*, *Port Scan*, and *Infiltration*—utilizing data from Kitsune and the CICIDS-2017 dataset.

Evaluation. To assess the effectiveness of the mimicked packets in evading the 6 classifiers, we employed a singular, highly illustrative metric from the Traffic Manipulator: the Malicious traffic Evasion Rate (MER). This metric is calculated using the formula: $MER = 1 - \left(N^{\mathrm{adv}}/N^{\mathrm{mal}}\right)$, where N^{adv} and N^{mal} represent the number of detected adversarial and detected malicious packet sequences, respectively. This equation captures the percentage of adversarial packet sequences that goes undetected as compared to the original malicious packet sequences that was detected, effectively measuring the evasion capability of the adversarial packet sequences.

We employed the two-sample Kolmogorov-Smirnov (K-S) test, a powerful non-parametric method used to determine the probabilistic differences between two data samples. Specifically, we compare the empirical cumulative distribution functions (eCDFs) of both the original malicious packets and the adversarial packets generated by our model. The K-S test calculates the maximum distance (K-S statistic, D) between these two eCDFs. This statistic measures the greatest deviation between the distribution of our adversarially modified packets and the original malicious packets (Hore et al. [2023]; Gretton et al. [2012]). A D value exceeding the critical threshold at a chosen significance level suggests a rejection of the null hypothesis—that is, the two samples are not derive from the same distribution. We demonstrated adversarial packets that were found to be out-of-distribution (OOD) at 95% significance level.

Additionally, we used Wireshark, a widely recognized network protocol analyzer to ensure that during the modifications were made, the immutable fields crucial for the packet's attack functionality were preserved. This visual verification via Wireshark confirms that our adversarial packets maintain their structural integrity and functionality, successfully mimicking genuine network traffic while evading detection (Due to limited space, results and analysis is in Appendix B.1).

4.2 Results and Analysis

Evasion Rate. We evaluated the evasion rates of the 6 attacks using DiffuPac and the baseline model, and compared the results in Table 2 (Due to space constraints, the other 4 attacks are demonstrated in Appendix B.2). As shown in Table 2, we conclude that DiffuPac is capable of generating adversarial packets that achieves comparable or even higher in evasion rate compared with the baseline. Importantly, DiffuPac achieves this high performance without relying on NIDS components, surrogate classifiers or specific insights into the NIDS's feature extraction methods, highlighting its effectiveness in realistic settings where attackers lack access to such insider information. The evasion performance varied significantly across different classifiers, reflecting the inherent variability in their robustness and detection capabilities. Flow-based NIDS exhibited greater resilience against our attacks compared to packet-based NIDS, likely due to DiffuPac's focus on packet-level modifications. Interestingly, the baseline model performed well with KitNet's AfterImage feature extractor, likely because it was trained directly on these features, enhancing its capability to modify them to evade detection. This specific alignment with NIDS features, while effective, does not typically reflect real-world attacker capabilities or constraints.

Table 3 : Percentage of successful adversarial samples found to be OOD.

Attack Type	Percentage of OOD Samples (%)
Botnet	48.72
MITM	29.34
Portscan	55.23
DDoS	78.14
Infiltration	58.06
Brute Force	69.67

Based on the results of all the attacks, we can conclude that ML classifier are more robust than DNNs to adversarial attacks. Traditional ML models, like DT and IF, have simpler and more interpretable decision boundaries, making them less susceptible to subtle adversarial modifications (Sauka et al. [2022]). In contrast, DNNs with their complex and high-dimensional decision boundaries are more easily misled by the nuanced modifications introduced by DiffuPac.

Statistical Difference (K-S test). The percentage of successful adversarial packets that were found to be OOD are demonstrated in Table 3. From the result, we can obeserve that Brute Force and DDoS attacks show the highest percentage of OOD samples. The grounded reasons for these kind of attacks are due to the inherent nature of their traffic patterns. The nature of Brute Force attacks involves numerous failed login attempts, leading to highly irregular sequence numbers (Javed and Paxson [2013]). Even after modifications by DiffuPac, these attempts will stand out due to their frequency and pattern, resulting in significant deviations in the eCDFs. The sheer volume and repetitive nature of DDoS traffic (Haseeb-ur rehman et al. [2023]) make it difficult to disguise effectively. Despite efforts of intelligent mutation process, the high volume and distinctive traffic patterns will cause substantial deviations in the eCDFs as well. On the other hand, the MITM show the lowest percentage of OOD samples. Adjusting TTL values is straightforward as they can be set to typical ranges seen in normal packets without significantly altering packet behavior, resulting in minimal deviation in the eCDFs.

5 Limitations

Our evaluations primarily focus on the evasion capabilities of adversarial packets and their ability to mimic legitimate network behavior, particularly in packet header fields. However, the impact of these modifications on the actual malicious functionality has not been thoroughly examined, raising questions about whether the packets retain their intended malicious effects undetected. Future studies will utilize honeypots and virtual network systems to better assess these aspects.

While DiffuPac effectively handles various attack types, its performance is inconsistent across different scenarios, notably DDoS and Brute Force. This variability is due to the distinct nature of each attack's traffic patterns, which may not be fully captured by the model, especially in highly repetitive or anomalous behaviors. To enhance the model's adaptability and generalizability, we plan to enrich the training dataset with a broader spectrum of real-world attack scenarios.

Moreover, the dual-use nature of adversarial generation models like DiffuPac presents significant ethical and legal challenges. While designed to improve security defenses, these technologies could be misused for malicious activities, such as facilitating cyber-attacks or disrupting services. To mitigate these risks, we have opted not to publicly release saved model checkpoints, aiming to prevent exploitation by malicious entities and ensure that our advancements in adversarial packet generation are used responsibly and within ethical bounds.

6 Conclusion and Future Directions

In this study, we introduced DiffuPac, an intelligent generative model that successfully generates adversarial packets capable of evading advanced NIDS while maintaining attack functionality. Unlike previous research, which often assumes attacker access to NIDS, DiffuPac operates under constraints of limited attacker knowledge, reflecting more realistic scenarios. Here, we present the in-depts insights obtained from this study: (a) DiffuPac uniquely combines normal and malicious packet sequences using contextual alignments, ensuring seamless integration into genuine traffic while employing these normal packet sequences to guide the denoising process. Through the performance evaluations with various NIDS, our model achieved an average improvement of approximately *8.83* percentage points in evasion rate compared to the baseline across all attack types. (b) Evasion rates varied notably across attack types, with DDoS and Brute Force attacks showing higher probabilistic differences due to their complex, repetitive nature. This highlights potential areas for further refinement in DiffuPac's approach to handling voluminous attacks. (c) Simpler ML models like DT and IF displayed surprising resilience due to their less complex decision boundaries, limiting the effectiveness of DiffuPac's modifications. In contrast, DNNs, with their intricate decision boundaries, were more vulnerable, underscoring the complexities inherent in designing robust adversarial tactics. (d) DiffuPac's capability to balance sophisticated attacks with operational stealth makes it especially suitable for environments where attackers lack comprehensive NIDS configurations, enhancing its utility in realistic defense testing.

As future studies, we will compare DiffuPac against a broader generative adversarial packets model and extend testing across more diverse NIDS to better understand its relative strengths and limitations. Next, we also learn that there is a notable lack of comprehensive evaluation concerning the functionality of adversarial packet attacks, particularly in assessing whether these packets retain their intended malicious capabilities after evading detection. We plan to integrate honeypots and virtually simulated network environments in future validations to test the practical effectiveness of DiffuPac's adversarial packets in real-world settings.

References

Statista. Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2023, with forecasts from 2022 to 2030. https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/, 2023. Accessed: 2024-4-20.

Tala Talaei Khoei and Naima Kaabouch. Machine learning: Models, challenges, and research directions. *Future Internet*, 15(10):332, 2023. doi: 10.3390/fi15100332.

Tala Talaei Khoei, Hadjar Ould Slimane, and Naima Kaabouch. Deep learning: Systematic review, models, challenges, and research directions. *Neural Computing and Applications*, 35(31):23103–23124, 2023. doi: 10.1007/s00521-023-08957-4.

Olakunle Ibitoye, Rana Abou Khamis, Ashraf Matrawy, and M. Omair Shafiq. The threat of adversarial attacks on machine learning in network security – A survey. *arXiv*, 2019. URL http://arxiv.org/abs/1911.02621.

Hassan Khazane, Mohammed Ridouani, Fatima Salahdine, and Naima Kaabouch. A holistic review of machine learning adversarial attacks in IoT networks. *Future Internet*, 16(1):32, 2024. doi: 10.3390/fi16010032.

Ivan Homoliak, Martin Teknøs, Martîn Ochoa, Dominik Breitenbacher, Saeid Hosseini, and Petr Hanacek. Improving network intrusion detection classifiers by non-payload-based exploit-independent obfuscations: An adversarial approach. *EAI Endorsed Transactions on Security and Safety*, 5(17), 12 2018. doi: 10.4108/eai.10-1-2019.156245.

Mohammad J. Hashemi, Greg Cusack, and Eric Keller. Towards evaluation of nidss in adversarial setting. In *Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks*, 2019.

Aditya Kuppa, Slawomir Grzonkowski, Muhammad Rizwan Asghar, and Nhien-An Le-Khac. Black box attacks on deep anomaly detectors. In *Proceedings of the 14th International Conference on Availability, Reliability and Security (ARES 2019)*, 2019.

Dongqi Han, Zhiliang Wang, Ying Zhong, Wenqi Chen, Jiahai Yang, Shuqiang Lu, Xingang Shi, and Xia Yin. Evaluating and improving adversarial robustness of machine learning-based network intrusion detectors. *IEEE Journal on Selected Areas in Communications*, 39(8):2632–2647, 2021. doi: 10.1109/JSAC.2021.3087242.

Yam Sharon, David Berend, Yang Liu, Asaf Shabtai, and Yuval Elovici. TANTRA: Timing-based adversarial network traffic reshaping attack. *IEEE Transactions on Information Forensics and Security*, 17:3225–3237, 2021. doi: 10.1109/TIFS.2022.3201377.

Soumyadeep Hore, Jalal Ghadermazi, Diwas Paudel, Ankit Shah, Tapas K. Das, and Nathaniel D. Bastian. Deep PackGen: A deep reinforcement learning framework for adversarial network packet generation. *arXiv*, 2023. URL https://arxiv.org/abs/2305.11039.

Zeeshan Ahmad, Adnan Shahid Khan, Cheah Wai Shiang, Johari Abdullah, and Farhan Ahmad. Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Transactions on Emerging Telecommunications Technologies*, 32(1):e4150, 2021. doi: https://doi.org/10.1002/ett.4150.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *Proceedings of the 2nd International Conference on Learning Representations (ICLR 2014)*, 2014.

Andrew McCarthy, Essam Ghadafi, Panagiotis Andriotis, and Phil Legg. Functionality-preserving adversarial machine learning for robust classification in cybersecurity and intrusion detection domains: A survey. *Journal of Cybersecurity and Privacy*, 2(1):154–190, 2022. doi: 10.3390/jcp2010010.

Kaichen Yang, Jianqing Liu, Chi Zhang, and Yuguang Fang. Adversarial examples against the deep learning based network intrusion detection systems. In *Proceedings of 2018 IEEE Military Communications Conference (MILCOM 2018)*, pages 559–564, 2018. doi: 10.1109/MILCOM.2018.8599759.

Ryan Sheatsley, Nicolas Papernot, Michael J. Weisman, Gunjan Verma, and Patrick McDaniel. Adversarial examples for network intrusion detection systems. *Journal of Computer Security*, 30 (5):727–752, 2022. doi: 10.3233/JCS-210094.

Ke He, Dan Dongseong Kim, and Muhammad Rizwan Asghar. Adversarial machine learning for network intrusion detection systems: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 25(1):538–566, 2023. doi: 10.1109/COMST.2022.3233793.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Proceedings of Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pages 6840–6851, 2020.

Xinjie Lin, Gang Xiong, Gaopeng Gou, Zhen Li, Junzheng Shi, and Jing Yu. Et-bert: A contextualized datagram representation with pre-training transformers for encrypted traffic classification. In *Proceedings of the ACM Web Conference 2022 (WWW 2022)*. ACM, 2022. doi: 10.1145/3485447.3512217.

Xuying Meng, Chungang Lin, Yequan Wang, and Yujun Zhang. Netgpt: Generative pretrained transformer for network traffic. *arXiv*, 2023. URL https://arxiv.org/abs/2304.09513.

Yonghui Wu, Mike Schuster, Z. Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason R. Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv*, 2016. URL https://arxiv.org/abs/1609.08144.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019)*, pages 4171–4186. ACL, 2019. doi: 10.18653/v1/N19-1423.

Chuanpu Fu, Qi Li, Meng Shen, and Ke Xu. Realtime robust malicious traffic detection via frequency domain analysis. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS 2021)*, pages 3431–3446. ACM, 2021. doi: 10.1145/3460120.3484585.

Jianbang Dai, Xiaolong Xu, and Fu Xiao. Glads: A global-local attention data selection model for multimodal multitask encrypted traffic classification of IoT. *Computer Networks*, 225:109652, 2023. doi: https://doi.org/10.1016/j.comnet.2023.109652.

Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-LM improves controllable text generation. In *Proceedings of Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, pages 4328–4343, 2022.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. DiffuSeq: Sequence to sequence text generation with diffusion models. In *Proceedings of the 11th International Conference on Learning Representations (ICLR 2023)*, 2023.

Alistair Letcher. On the impossibility of global convergence in multi-loss optimization. *arXiv*, 2021. URL https://arxiv.org/abs/2005.12649.

Yisroel Mirsky, Tomer Doitshman, Yuval Elovici, and Asaf Shabtai. Kitsune: An ensemble of autoencoders for online network intrusion detection. In *Proceedings 2018 Network and Distributed System Security Symposium (NDSS 2018)*, 2018.

Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP 2018)*, pages 108–116. INSTICC, SciTePress, 2018. doi: 10.5220/0006639801080116.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv*, 2024. URL https://arxiv.org/abs/2401.08281.

Gerard Draper-Gil, Arash Habibi Lashkari, Mohammad Saiful Islam Mamun, and Ali A. Ghorbani. Characterization of encrypted and vpn traffic using time-related features. In *International Conference on Information Systems Security and Privacy (ICISSP 2016)*, 2016. doi: 10.5220/0005740704070414.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Kudzai Sauka, Gun-Yoo Shin, Dong-Wook Kim, and Myung-Mook Han. Adversarial robust and explainable network intrusion detection systems based on deep learning. *Applied Sciences*, 12(13), 2022. doi: 10.3390/app12136451.

Mobin Javed and Vern Paxson. Detecting stealthy, distributed SSH brute-forcing. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer and Communications Security (CCS 2013)*, pages 85–96, New York, NY, USA, 2013. ACM. doi: 10.1145/2508859.2516719.

Rana M. Abdul Haseeb-ur rehman, Azana Hafizah Mohd Aman, Mohammad Kamrul Hasan, Khairul Akram Zainol Ariffin, Abdallah Namoun, Ali Tufail, and Ki-Hyung Kim. High-speed network ddos attack detection: A survey. *Sensors*, 23(15), 2023. doi: 10.3390/s23156850.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 6000–6010, 2017.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the key contributions of our paper, including the development of DiffuPac, the integration of a pre-trained BERT model with diffusion model, and the novel approach to generating adversarial packets. These claims are substantiated by both theoretical analysis and experimental results presented in the paper. The scope of our research, including the focus on real-world constraints and the effectiveness of the proposed model against advanced NIDS, is accurately reflected in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have included the limitations in the main paper.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The proofs for all theorems and formulas are clearly provided both in the main paper and supplemental material.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides a detailed explanation of the model architecture and the experimental setups. This includes a thorough description of the pre-trained BERT model integration, the diffusion model, and the specific configurations used in the experiments. Additionally, all hyperparameters, dataset details, and evaluation metrics are disclosed to enable replication of the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Due to ethical concerns, we have decided not to publicly share the data and code. Additionally, much of the code is based on existing codebases from other sources. However, we have provided a full and detailed explanation of the model architecture and its implementation in the paper and supplemental material. This includes comprehensive descriptions of the experimental setups, hyperparameters, and step-by-step instructions to facilitate reproducibility.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have provided the full detailed explanation regarding the training process and the experimental details both in the main paper and supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The experiments conducted in this paper primarily focus on evaluating the success of generating adversarial packets. Metrics such as evasion rate, packet field analysis, and probabilistic differences between original and adversarial packets are used to assess the effectiveness of the proposed method.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We have detailed the types of computational resources used and the time of execution for BERT and diffusion model in the supplemental materials.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have read the Code Of Ethics and strictly follow it.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed the positive societal impacts and negative societal impacts in the main paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Due to ethical concerns of our paper, we have implemented stringent measures.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: All existing assets used in this paper, including code, data, and models, have been properly credited to their original creators. We have cited the original papers that produced these assets and included URLs where applicable.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces a new model, DiffuPac, which integrates a pre-trained BERT with a diffusion model to generate adversarial packets. Detailed documentation for DiffuPac is provided, including descriptions of the model architecture, training process, implementation details, and usage guidelines. This documentation is included in the main paper and supplemental materials.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

A Model Architecture

A.1 BERT

The core of DiffuPac is the BERT model, specifically adapted to serve as the denoising component in the diffusion process. The architecture consists of 12 bidirectional Transformer blocks and each block incorporates 12 attention heads within its self-attention layers (Vaswani et al. [2017]). These layers are crucial for capturing the intricate and implicit relationships between traffic bytes. Each input token has a dimension of 768, and the model can process sequences with up to 512 tokens. This setup allows the BERT model to effectively handle extensive sequences of traffic data, making it well-suited for the task of denoising during the diffusion process.

To implement of our pre-trained BERT model, we utilized the codebase from the UER (Universal Encoder Representations)-py project. UER-py is a versatile toolkit designed for pre-training and fine-tuning various NLP models, including BERT. In UER-py, the MLM is designed to predict masked words within a sentence. We adapted this approach to our masked unidirectional flow model, which focuses on predicting masked tokens within network traffic sequences. Instead of textual tokens, our model deals with traffic units from src-to-dst and dst-to-src sequences. We modified the MLM implementation to handle these traffic tokens, ensuring that each token is masked with a probability of 15%. Our model then predicts the original token from the masked sequence, capturing semantic patterns within bidirectional network flows.

The Next Sentence Prediction (NSP) task in UER-py involves prediction step to determine whether one sentence follows another. We transformed this concept into the same sequence-origin prediction task to determine the origin of directional network traffic. We adapted the binary classifier used in NSP to classify whether packets are originated from src-to-dst or dst-to-src. By pairing packets and predicting the directional flow, our model learns the contextual relationships and patterns in network traffic, similar to how NSP learns sentence relationships.

Table 4 lists the parameters are set to run our pre-trained BERT model. The remaining parameters were set to their default values.

Table 4: Parameters used for pre-training the BERT model.

Parameter	**Value**
Embedding Size	768
Feedforward Size	3072
Hidden Size	768
Hidden Activation Function	GELU
Number of Attention Heads	12
Number of Transformer Layers	12
Max Sequence Length	512
Dropout Rate	0.1
Batch Size	64
Total Steps	500000
Learning Rate	1e-4

A.2 Fine-tuning with Diffusion Model

For the fine-tuning phase of our model, we leveraged the DiffuSeq codebase (DiffuSeq's code). The DiffuSeq framework, originally designed for text data, required several significant adaptations to handle network traffic data and to meet the objectives of our diffusion model for adversarial packet generation. The primary adaptation to the DiffuSeq codebase involves the data preparation phase, specifically the concatenation strategy used to blend normal and malicious packet sequences. This adaptation is crucial for the model to learn how to generate adversarial packets that can mimic normal packets and evade detection by NIDS.

In our concatenation strategy, we identified pairs of normal and malicious packet sequences that exhibit strong contextual alignments unlike DiffuSeq, which, randomly merge the text sequences.

This means that the patterns and behaviors within these sequences are similar, ensuring that the concatenated sequences are blended seamlessly. We concatenated these contextually aligned normal and malicious packet sequences to form a new sequence, $\mathbf{S}^{\text{ben}\oplus\text{mal}}$. The strategy leverages the inherent patterns and behaviors within normal packet sequences, which is crucial for the mimicking process of malicious packet sequences.

We utilized the pre-trained BERT embeddings in our concatenation strategy. BERT is designed to capture deep semantic relationships within sequences. This design makes BERT making it highly effective for analyzing complex patterns in network traffic. By employing the pre-trained embeddings, DiffuPac can accurately transform discrete packet data into continuous feature spaces, capturing the intricate dependencies between packets. This transformation is critical for the subsequent similarity analysis and pairing process. The pre-trained embeddings offer a rich representation of the packet sequences, so that the model can discern subtle contextual similarities between normal and malicious packet sequences. The detailed of the concatenation strategy is demonstrated in Algorithm 1.

Here, we set the similarity threshold, ϵ to 0.9. This is a strategic decision aimed at balancing the need for high contextual alignment with the practicalities of ensuring a sufficient number of matched pairs. Setting ϵ to 0.9 indicates that we are looking for pairs of packet sequences that are highly similar, capturing nearly all the essential contextual information shared between normal and malicious packet sequences. This high threshold ensures that the concatenated sequences are blended seamlessly, maintaining the realistic traffic patterns and behaviors required to evade detection by NIDS. From a practical standpoint, a threshold of 0.9 is chosen because it strikes an optimal balance between precision and recall. A higher threshold (closer to 1.0) would result in fewer matched pairs, as only the most similar sequences would be selected. This could limit the model's ability to generate a diverse set of adversarial packets. Conversely, a lower threshold would increase the number of matched pairs but might include sequences that are not contextually well-aligned, reducing the effectiveness of the adversarial packets in mimicking normal traffic.

During the reverse process, we introduced a guidance mechanism using normal packet sequences. This mechanism treats the concatenated sequences of normal and noise-added malicious packets as a unified entities, allowing the model to denoise the malicious packets in the context of the normal packet sequences. We recalibrated the variational lower bound L_{VLB} to emphasize the integration of normal sequence guidance. The objective focuses on accurately reconstructing the initial state, $\mathbf{z}_0$ from the noised states, so that the malicious components are seamlessly integrated into the normal packet sequences pattern.

Table 5 lists the parameters used to run our diffusion model. The other parameters are set according to the default values of DiffuSeq. All training, experimentation, and sampling processes are executed on a single NVIDIA AD102 (GeForce RTX 4090) GPU. The training process for BERT required approximately 23 hours, while fine-tuning with the diffusion model took roughly 10 hours.

Table 5: Parameters used for the diffusion model.

Parameter	**Value**
Diffusion Steps	2000
Learning Rate	1e-4
Learning Steps	50000
Seed	102
Noise Schedule	sqrt
Batch Size	64
Microbatch	64
Sequence Length	128
Hidden Time Dimension	128
Hidden Dimension	128
Schedule Sampler	uniform

Algorithm 1 Finding contextually relevant packet sequences.

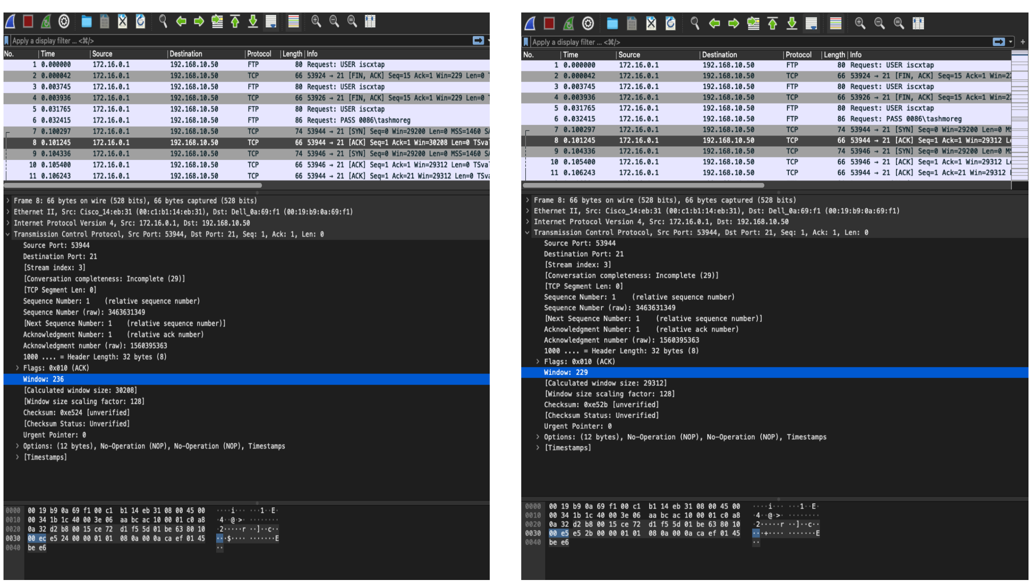
Require: Normal packet sequences $\mathbf{S}^{\text{ben}}$, Malicious packet sequences $\mathbf{S}^{\text{mal}}$, Model parameters θ, Similarity threshold ϵ

Ensure: Matched pairs of normal and malicious packet sequences $(\mathbf{S}_i^{\text{ben}}, \mathbf{S}_j^{\text{mal}})$

```
Initialize: Load pre-trained BERT model with parameters θ
Embed Normal Sequences:
for each sequence S_i^ben in S^ben do
    Compute embedding E_i^ben ← EMB(S_i^ben; θ)
end for
Embed Malicious Sequences:
for each sequence S_j^mal in S^mal do
    Compute embedding E_j^mal ← EMB(S_j^mal; θ)
end for
Find Contextually Aligned Pairs:
for each embedding E_j^mal in E^mal do
    Initialize best match b_j ← None
    Initialize highest similarity σ_max ← −1
    for each embedding E_i^ben in E^ben do
        Compute similarity σ_ij ← cosine_similarity(E_j^mal, E_i^ben)
        if σ_ij > σ_max then
            Update best match b_j ← S_i^ben
            Update highest similarity σ_max ← σ_ij
        end if
    end for
    if σ_max > ϵ then
        Add pair (b_j, S_j^mal) to matched pairs
    end if
end for
Return Matched Pairs: Return all matched pairs (S_i^ben, S_j^mal)
```

B Experimental Results

B.1 Wireshark Results

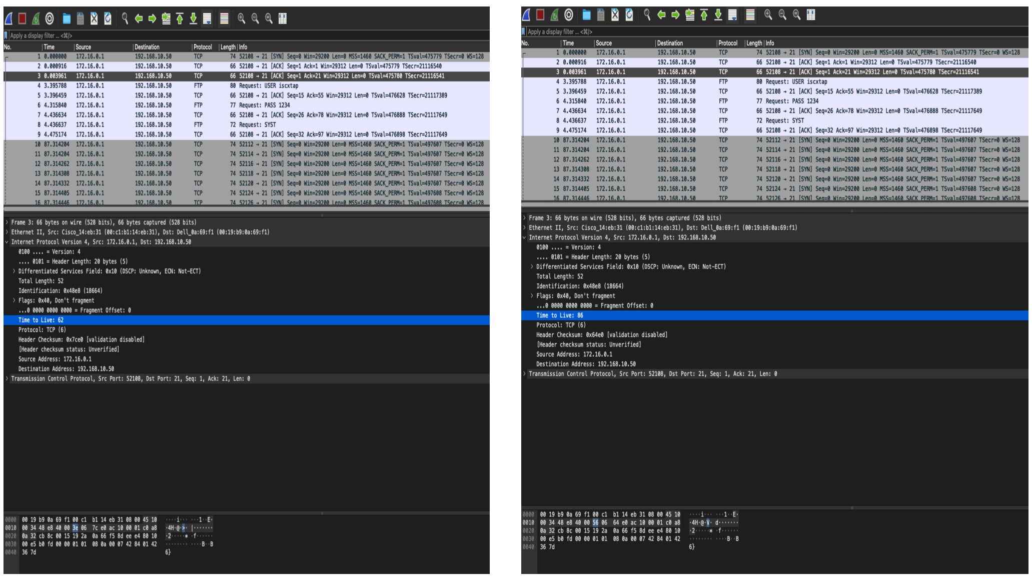


(a) Botnet attack before in Wireshark. (b) Botnet attack after in Wireshark.

Figure 2: Comparison of Botnet attack before and after in Wireshark.

To demonstrate that our model intelligently alters specific fields to seamlessly blend malicious packets into normal traffic, we analyzed the packets using Wireshark. Our analysis covered 6 attack types: *MITM*, *Botnet*, *Brute Force*, *DDoS*, *Port Scan*, and *Infiltration*. The results showed that different specific fields were modified in the packets corresponding to each type of attack. This variation in field alteration confirms that our model adapts its modifications according to the nature of the attack, enhancing the stealthiness of the malicious packets. In Figure 2, we can see that there is a change in the window size of a Botnet packet. Botnet attack can be described as a network of infected devices (bots) controlled by an attacker to perform various malicious activities. Botnet traffic often shows abnormal patterns in window size due to automated control and data bursts. Here, we can conclude that DiffuPac made the intelligent decision by itself in modifying the window size to fall within typical user traffic patterns. This helps to camouflage the botnet activity.

MITM, on the other hand, demonstrated attacks whereby an attacker intercepts communication between two parties without their knowledge, often to eavesdrop or alter the data being sent. Abnormal TTL values (time-to-live field) can reveal this kind of interception and redirection. Here, once again, our model intelligently adjusts TTL value to match the expected values in aiding the masking of MITM presence. As shown in Figure 3, DiffuPac adjusts the TTL values to effectively conceal signs of interception. This ensures that the adversarial packets blend seamlessly with normal data exchanges, mimicking legitimate communication between parties.



(a) MITM attack before in Wireshark.

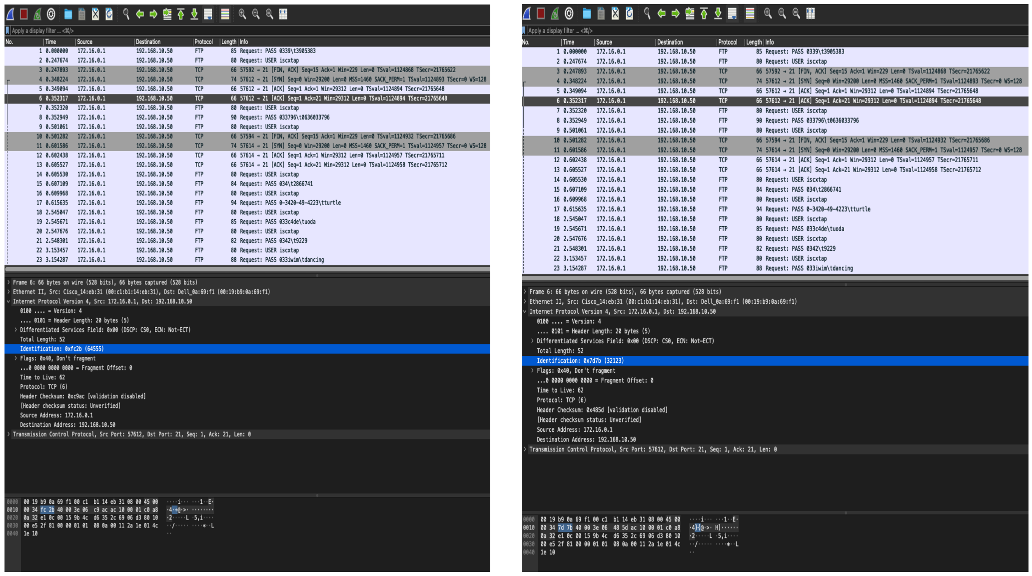
(b) MITM attack after in Wireshark.

Figure 3: Comparison of MITM attack before and after in Wireshark.

DDoS attacks involve a large number of packets with similar characteristics, often reusing or predictably incrementing the IP Identification (ID) field. In Figure 4, it can be shown that DiffuPac addresses this by randomizing the IP IDs to prevent detection. When we analyzed the traffic using Wireshark, we observed varied and less predictable IP IDs in the modified packets, aligning with normal traffic patterns. This randomness helps avoid forming detectable patterns and evades detection systems that rely on identifying repetitive IP ID sequences. Thus, DiffuPac successfully masks the attack's presence while preserving the attack functionality.

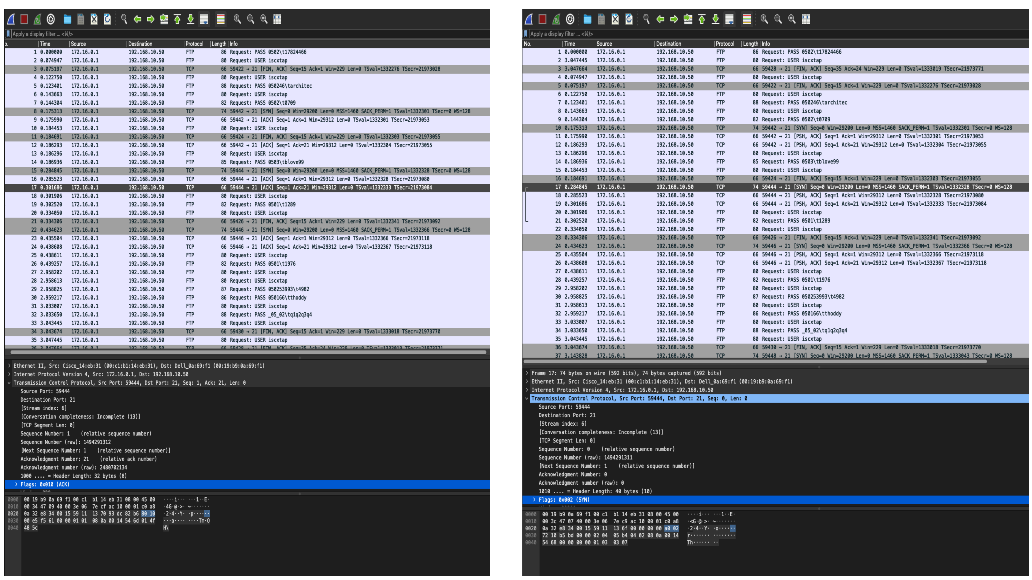
PortScan often involve specific flags, such as SYN, to probe for open ports. This behavior is distinct from normal traffic, which uses a variety of flags. In Figure 5, we found that DiffuPac diversifies the use of flags to include SYN, ACK, and FIN, similar to normal connections. By normalizing the traffic in this way, DiffuPac helps disguise scanning activity. For example, instead of sending many SYN packets, it includes a mix of SYN, ACK, and FIN packets, which helps blend the scan into typical network behavior and evade detection.

Brute force attacks generate numerous login attempts, leading to irregular sequence numbers due to the repeated connections. By analyzing the traffic with Wireshark, we observed that DiffuPac intelligently normalizes sequence numbers to mimic regular user connections. Normal user login



(a) DDoS attack before in Wireshark. (b) DDoS attack after in Wireshark.

Figure 4: Comparison of DDoS attack before and after in Wireshark.

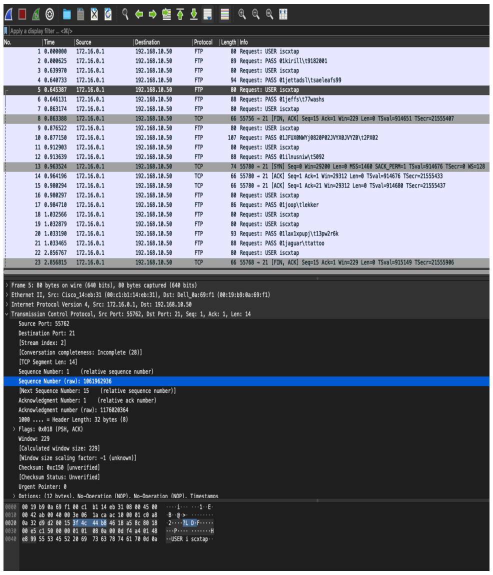


(a) PortScan attack before in Wireshark. (b) PortScan attack after in Wireshark.

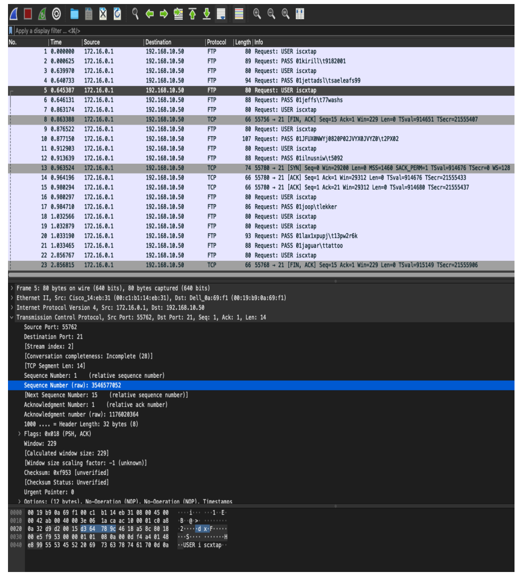
Figure 5: Comparison of PortScan attack before and after in Wireshark.

attempts typically show sequentially increasing sequence numbers. In Figure 6, it shows that DiffuPac adjusts these numbers to follow a similar pattern, thereby camouflaging the brute force activity. This adjustment helps blend the attack packets seamlessly into benign traffic, making it harder for NIDS to spot the abnormality.

Infiltration attacks involve moving laterally within a network, often generating unusual acknowledgment numbers as the attacker accesses various systems. In Figure 7, it demonstrated that DiffuPac intelligently modifies acknowledgment numbers to follow expected sequences. In normal communication, acknowledgment numbers increment predictably based on the received data. DiffuPac ensures that infiltration attempts have acknowledgment numbers that follow these normal patterns, effectively masking the lateral movement within the network and blending the attack traffic with normal traffic, making it difficult for detection systems to identify the anomaly.

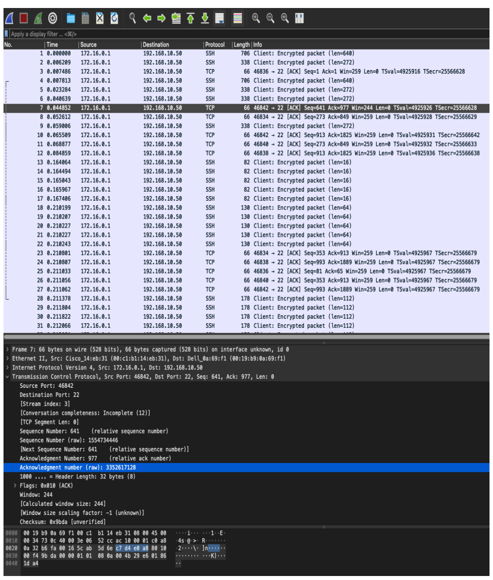


(a) BruteForce attack before in Wireshark.

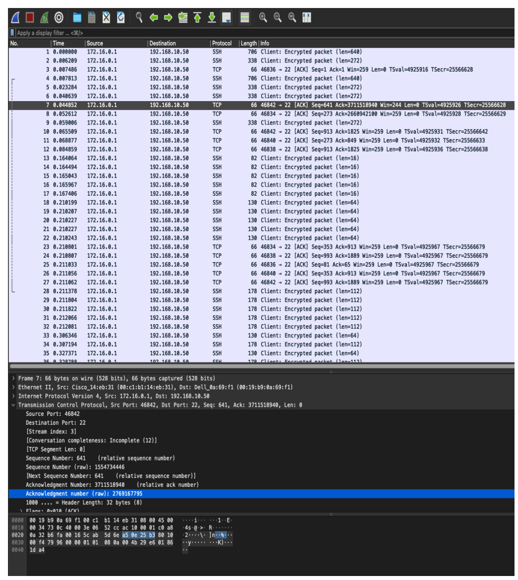


(b) BruteForce attack after in Wireshark.

Figure 6: Comparison of bruteForce attack before and after in Wireshark.



(a) Infiltration attack before in Wireshark.



(b) Infiltration attack after in Wireshark.

Figure 7: Comparison of infiltration attack before and after in Wireshark.

DiffuPac demonstrates its ability to modify specific fields intelligently in the packet headers of different attack types, such as the window size for botnet traffic and TTL for MITM, to blend malicious packets into normal packets. This capability is consistently proven through detailed packet analysis with Wireshark. Indeed, the detailed packets analysis shows that critical fields defining packet identity and practicality (e.g., IP addresses, ports, payload) remain intact after modification, ensuring the functionality of the attack while evading detection by sophisticated NIDS. The comparisons for each attack type before and after the modificatinos further highlight DiffuPac's effectiveness in maintaining attack practicality while enhancing stealth.

B.2 Evasion Rate

We evaluated the evasion rates of the other 4 attacks—DDoS, PortScan, Brute Force, and Infiltration—using DiffuPac and the baseline model, with the detailed results summarized in Table 6. According to the results, we can observed that the trends in the 4 attacks are consistent with the 2 attacks shown in Table 2 in the main paper.

Table 6: Comparative analysis of attack detection and evasion rates of the other 4 attacks.

(c) DDoS

Feature Extractor	Classifier	Detection			Evasive (*MER*)	
		P	R	F1	Baseline	**Ours**
CIC FLowMeter	KitNET	0.93	0.90	0.91	32.71%	**51.96%**
	DT	0.71	0.79	0.72	38.67%	**46.23%**
	IF	0.99	0.90	0.92	25.95%	**50.35%**
	MLP	0.72	0.71	0.73	52.03%	**69.87%**
	SVM	0.73	0.76	0.78	42.09%	**54.19%**
	LR	0.73	0.78	0.76	48.45%	**59.82%**
AfterImage	KitNET	0.97	0.95	0.96	59.84%	**62.31%**
	DT	0.78	0.92	0.85	**72.88%**	59.47%
	IF	0.86	0.84	0.88	22.63%	**48.91%**
	MLP	0.91	0.94	0.92	55.45%	**63.25%**
	SVM	0.99	0.90	0.94	**78.51%**	68.93%
	LR	0.94	0.97	0.91	**71.68%**	56.66%

(d) PortScan

Feature Extractor	Classifier	Detection			Evasive (*MER*)	
		P	R	F1	Baseline	**Ours**
CIC FLowMeter	KitNET	0.95	0.91	0.94	32.31%	**54.96%**
	DT	0.78	0.72	0.71	34.77%	**59.83%**
	IF	0.99	0.90	0.94	21.61%	**43.01%**
	MLP	0.76	0.71	0.74	60.33%	**68.12%**
	SVM	0.73	0.76	0.78	**67.40%**	60.10%
	LR	0.78	0.82	0.79	43.81%	**54.93%**
AfterImage	KitNET	0.94	0.96	0.93	**69.26%**	55.87%
	DT	0.77	0.91	0.82	69.78%	**70.47%**
	IF	0.84	0.91	0.89	18.38%	**48.30%**
	MLP	0.96	0.94	0.97	49.73%	**64.02%**
	SVM	0.97	0.96	0.95	**76.40%**	75.81%
	LR	0.95	0.92	0.90	67.29%	**73.94%**

(e) BruteForce

Feature Extractor	Classifier	Detection			Evasive (*MER*)	
		P	R	F1	Baseline	**Ours**
CIC FLowMeter	KitNET	0.97	0.94	0.96	36.71%	**49.24%**
	DT	0.74	0.77	0.71	39.67%	**41.89%**
	IF	0.91	0.90	0.94	25.95%	**38.02%**
	MLP	0.88	0.85	0.83	50.63%	**65.81%**
	SVM	0.80	0.78	0.70	46.09%	**54.19%**
	LR	0.79	0.75	0.77	47.45%	**60.62%**
AfterImage	KitNET	0.96	0.93	0.95	**81.84%**	69.31%
	DT	0.82	0.79	0.78	41.07%	**55.07%**
	IF	0.88	0.98	0.92	24.43%	**39.71%**
	MLP	0.91	0.94	0.92	59.69%	**68.25%**
	SVM	0.94	0.92	0.96	**68.51%**	58.93%
	LR	0.97	0.93	0.94	61.68%	**64.06%**

(f) Infiltration

Feature Extractor	Classifier	Detection			Evasive (*MER*)	
		P	R	F1	Baseline	**Ours**
CIC FLowMeter	KitNET	0.89	0.90	0.95	39.91%	**43.86%**
	DT	0.72	0.77	0.78	30.32%	**61.46%**
	IF	0.99	0.90	0.94	25.95%	**43.19%**
	MLP	0.76	0.71	0.74	51.43%	**64.63%**
	SVM	0.73	0.76	0.78	**61.09%**	54.59%
	LR	0.71	0.79	0.76	41.28%	**59.17%**
AfterImage	KitNET	0.92	0.98	0.97	**73.44%**	51.26%
	DT	0.77	0.91	0.82	**69.23%**	62.54%
	IF	0.84	0.91	0.89	23.61%	**42.59%**
	MLP	0.96	0.94	0.97	50.65%	**72.25%**
	SVM	0.94	0.96	0.98	**72.48%**	64.73%
	LR	0.93	0.92	0.95	66.04%	**68.37%**