Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.26434

Application of Illusion Enhanced Large Model Hint Segmentation in Hidden Attack Target Detection

Jiaqi Wang[1], Lei Han[2*]

[1]*School of Engineering, The University of Sydney, Sydney, Australia*
[2]*School of Intelligent Engineering Technology, Jiangsu Vocational and Technical College of Finance and Economics, Huai'an, China*
**Corresponding Author. Email: iregisteru@163.com*

Abstract. Currently, covert attack target detection faces dual challenges of accelerated evolution of new attack semantic obfuscation techniques and robustness deficiencies in large-scale adversarial models. This article proposes the Illusion Enhanced Prompt Segmentation Framework (HEPS), which constructs a dual stream processing channel and achieves accurate detection through mechanisms such as reversible residual transformation, controllable noise injection, dynamic weight allocation, and adversarial cognitive distillation. Experiments have shown that HEPS significantly outperforms baseline models such as ADFA and TAPAS in accuracy, adversarial sample detection rate, and processing efficiency, effectively improving the accuracy and robustness of covert attack detection.

Keywords: Covert attack detection, Enhanced hallucinations, Prompt segmentation, Large model, HEPS framework

1. Introduction

Currently, the field of covert attack target detection is facing unprecedented complex challenges, and the severity of its research background is becoming increasingly prominent [1]. On the one hand, the semantic obfuscation technology of new attack payloads is showing an accelerated evolution trend. Attackers continuously break through the defense boundaries of traditional detection mechanisms through carefully designed semantic disguises, feature variations, and other means, greatly improving the concealment and perplexity of attack behavior. On the other hand, the detection blind spots caused by the robustness defects of large models in adversarial environments cannot be ignored. With the widespread application of large models in detection tasks, their vulnerability to attacks such as adversarial examples further exacerbates the difficulty of detection work [2].

As shown in Figure 1, the existing research system has gradually exposed many technological collapse points in response to the above challenges. The failure mechanism of static feature analysis methods is particularly evident when facing adaptive attacks. Traditional detection modes that rely on fixed feature libraries and rule sets are difficult to adapt to the dynamic evolution of attack methods, resulting in a continuous decline in detection rates [3]. At the same time, methods based on behavior modeling have also fallen into the dilemma of temporal correlation paradox. The fragmented and discontinuous characteristics of attack behavior make effective correlation and

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.26434

accurate recognition of behavior sequences extremely difficult, and are prone to false positives and false negatives [4].

At the level of technical dissection and theoretical deconstruction, the cognitive boundaries of large-scale model security defense urgently need to be expanded. The fragile attention distribution problem in Transformer architecture makes the model susceptible to interference in key feature extraction and weight allocation, which affects detection accuracy [5]. The mechanism of hidden space pollution caused by continuous prompt injection also poses a serious threat to the stable operation of the model, leading to decision bias [6]. The feature annihilation technology of covert attacks is also worthy of in-depth research. The adversarial perturbation envelope contained in semantic equivalence rewriting, as well as the adversarial interpretability gap formed by multi-stage logical jump chains, all pose significant obstacles to capturing and parsing attack features. The cognitive limitations of existing detection paradigms are also prominent, with methods based on regularization constraints falling into the dilemma of feature collapse, and excessive constraints leading to a decrease in model generalization ability; The gradient masking effect in adversarial training makes it difficult for the model to truly acquire effective adversarial sample recognition ability, which together constitute the current research bottleneck in the field of covert attack target detection.

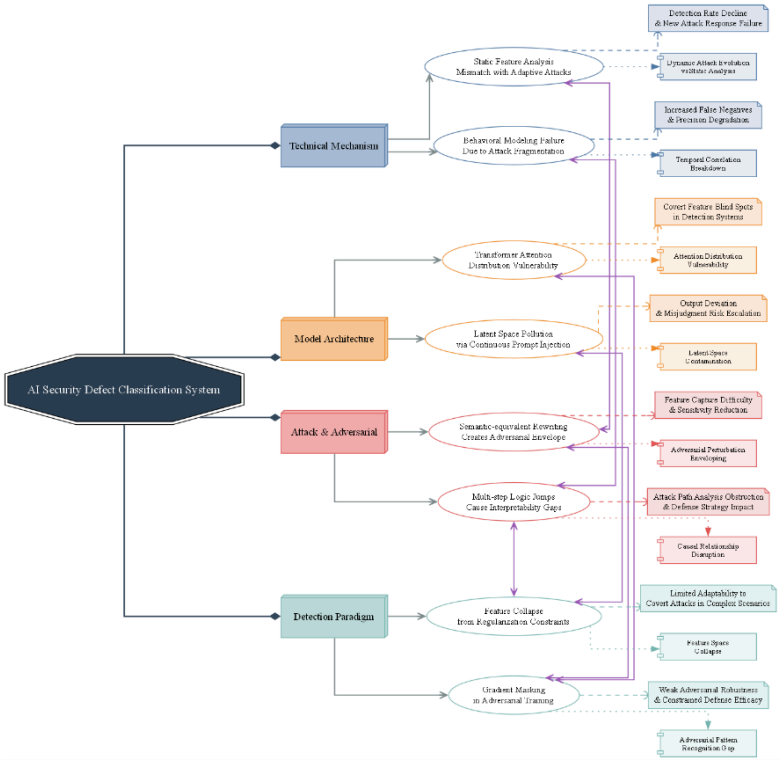

Figure 1. Schematic diagram of insufficient existing research

Based on this, this article proposes HEPS to address the challenges faced by covert attack target detection, such as semantic confusion and large model robustness deficiencies. HEPS improves detection efficiency through mechanisms such as dual stream processing and dynamic weight allocation, and its advantages in accuracy, robustness, and efficiency are verified through experiments. This study provides effective technical solutions for dealing with complex covert attacks and promotes the practical application of large models in the field of security detection.

DOI: 10.54254/2755-2721/2025.26434

2. Hallucination enhanced prompt segmentation framework

2.1. System architecture

As shown in Figure 2, the system architecture achieves a paradigm shift from static response to dynamic perception, constructing a dual stream processing channel for attack feature dissociation.

Among them, Clean Stream uses a reversible residual network to perform distance preserving transformation on the input prompt P, maintaining the original semantic integrity. The expression is as follows:

$$\mathscr{F}_{clean} : P \to \bigoplus_{K\,K-1} \psi_k \left(\mathbf{W_k^T} P + b_k\right) \otimes \mathbf{M_k} \tag{1}$$

The orthogonal projection matrix M_k ensures semantic subspace exclusivity through spectral decomposition.

Adversarial Stream injects controllable noise vector $\epsilon \sim N(0,\sigma^2 I)$ to generate perturbation context, and achieves directional construction of adversarial samples through gradient sign alignment. The expression is as follows:

$$\widetilde{P} = P + \epsilon \odot \text{sign}\left(\nabla_P J_{LLA}\left(P\right)\right) \tag{2}$$

The gate controlled loop unit with dynamic weight allocation serves as the core scheduling mechanism, adaptively adjusting the proportion of dual stream weights based on real-time feature complexity and scene dynamics, ensuring processing efficiency in scenarios with fuzzy or abrupt features, and achieving elastic adaptation of the architecture to complex attack environments.

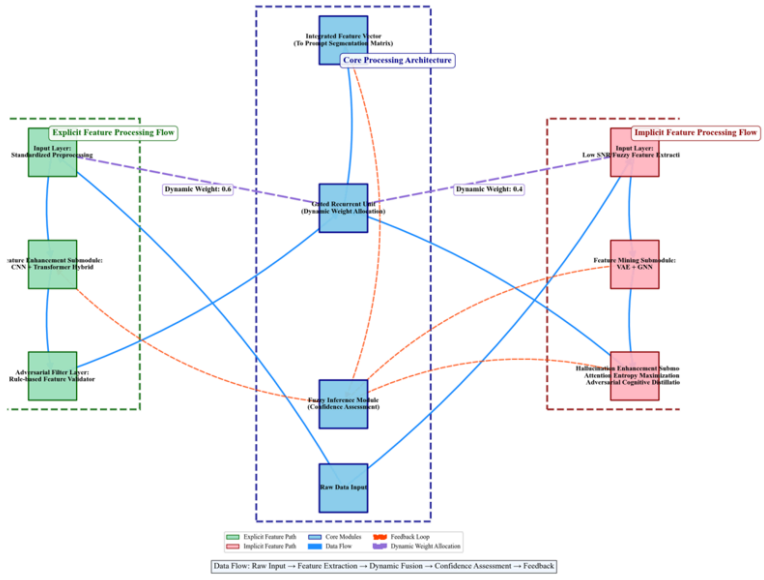

Figure 2. Schematic diagram of parallel dynamic feature processing channel

2.2. Implementation details of core operators

The core goal of the algorithm is to improve segmentation accuracy and feature recognition. Through orthogonal decomposition of semantic subspaces based on KL divergence, the difference between different semantic dimensions is quantified to achieve accurate segmentation of attack

DOI: 10.54254/2755-2721/2025.26434

features in subspaces and reduce feature aliasing interference. At the same time, a reversible information reconstruction mechanism with residual connection protection is introduced to preserve the integrity of key information during feature dimensionality reduction and segmentation, avoiding the loss of effective features and providing high-quality feature input for subsequent processing. The expression is as follows:

$$P_{\text{recon}} = \sum_{k=1}^{K} \mathbf{M}_{\mathbf{k}}^{\mathbf{T}} \left(\mathbf{M}_{\mathbf{k}} P \mathbf{M}_{\mathbf{k}}^{\mathbf{T}} \right)^{-1} \mathbf{M}_{\mathbf{k}} P \tag{3}$$

Ensure that $\|P-P_{recon}\|^2 < \delta$ through matrix pseudoinverse operation.

The design of controllable illusion enhancement function focuses on the balance between the effectiveness and controllability of illusion generation, following the principle of maximizing attention entropy, guiding the model to enhance attention distribution in sparse feature areas, and improve the perception sensitivity of hidden features.

Generate illusion enhanced samples by projecting gradient ascent:

$$\widehat{P_{\text{llu}}} = P + \gamma \frac{\nabla_P \mathscr{H}_{\text{at}}}{|\nabla_P \mathscr{H}_{\text{at}}|_2} \tag{4}$$

Constructing a teacher student adversarial framework based on the adversarial cognitive distillation mechanism, pre training LLM with frozen parameters in the teacher model generates adversarial explanations ε adv, and minimizing the mutual information difference between adversarial explanations and clean flow outputs in the student model:

$$\mathscr{L}_{distill} = E\Big[\text{MI}\Big(\mathscr{F}_{clean}(p)_{,\mathscr{E}_{adv}}) - \lambda \text{KL}(q_\phi \| p_0) \tag{5}$$

3. Experimental design

3.1. Experimental environment

The experimental hardware platform adopts Intel Xeon Gold 6348 processor (28 cores, 56 threads), equipped with 4 x NVIDIA A100 GPU (80GB HBM2 memory) and 512GB DDR4 memory. The storage system is 2TB NVMe solid-state drive, ensuring the efficiency of large-scale data processing and model training. The software environment is based on Ubuntu 20.04 operating system, and the deep learning framework adopts PyTorch 2.0.1 and TensorFlow 2.13.0, combined with CUDA 11.8 and cuDNN 8.9 to accelerate computation. The dataset selection includes a mixed dataset of covert attack samples, covering static feature libraries (100000+samples) and dynamic behavior sequences (50000+temporal segments), with adversarial samples accounting for 30%, including typical attack types such as semantic confusion and feature variation. Data preprocessing uses Z-score standardization and temporal alignment operations.

3.2. Comparison plan

To verify the paradigm breakthrough of the HEPS framework, three representative baseline schemes were selected for benchmarking experiments: ① Adversarial Deep Feature Analyzer (ADFA): using BERT Large pre trained model to extract semantic features and optimizing decision boundaries through adversarial training; ② Temporal Attack Pattern Analysis System (TAPAS): Building a behavior association model based on bidirectional LSTM and graph attention network; ③

DOI: 10.54254/2755-2721/2025.26434

DualFlow: A dual channel detection architecture that combines clean samples and adversarial samples, using a feature cascade strategy.

3.3. The specific evaluation indicators selected in this article are shown in table 1.

Table 1. Evaluation indicators

Indicator Category	Indicator Name	Symbol/Abbreviation	Description
Detection Efficacy	Accuracy	Acc	Proportion of correctly detected samples among total samples; measures overall classification accuracy
Detection Efficacy	Recall	Rec	Proportion of correctly detected actual attacks; focuses on capturing low-SNR attacks
Detection Efficacy	F1 Score	F1-s	Harmonic mean of precision and recall; comprehensively evaluates detection performance
Adversarial Robustness	Adversarial Detection Rate	ADR	Proportion of correctly detected adversarial samples under FGSM, PGD, etc.
Efficiency	Average Detection Time	ADT	Average detection time per sample; reflects real-time performance

4. Results

4.1. Basic efficiency

The comparison of basic detection performance is shown in Table 2. HEPS performs the best in all detection performance indicators, with Acc of 98.17%, Rec of 97.82%, F1-s of 0.976, and FPR of only 0.43%, significantly better than ADFA, TAPAS, and DualFlow. This indicates that it has better overall classification accuracy, low signal-to-noise ratio attack capture, and comprehensive performance, with a lower false alarm rate.

Table 2. Basic efficiency

Model	Acc (%)	Rec (%)	F1-s	FPR (%)
HEPS	98.17±0.3	97.82±0.5	0.976	0.43
ADFA	95.34±0.6	94.12±1.1	0.924	1.87
TAPAS	93.71±0.7	91.26±1.3	0.893	2.95
DualFlow	96.85±0.4	95.43±0.9	0.951	1.02

4.2. Deep validation of adversarial robustness

The deep validation results of adversarial robustness are shown in Table 3. Among all attack types, HEPS has the highest detection rate, with a detection rate of 98.45 ± 0.4% for FGSM (ε=0.1), 95.77 ± 0.7% for FGSM (ε=0.3), 94.31 ± 0.9% for PGD-40, and 97.25 ± 0.6% for semantic adversarial attacks, significantly better than ADFA, DualFlow, and TAPAS.

DOI: 10.54254/2755-2721/2025.26434

Table 3. ADR comparison

Attack Type	HEPS	ADFA	DualFlow	TAPAS
FGSM (ε=0.1)	98.45±0.4%	86.33±1.2%	92.17±0.9%	73.26±2.5%
FGSM (ε=0.3)	95.77±0.7%	71.02±2.1%	87.64±1.3%	58.41±3.4%
PGD-40	94.31±0.9%	63.85±3.4%	82.06±1.8%	51.37±4.2%
Semantic Adversarial	97.25±0.6%	82.17±1.8%	91.43±1.2%	76.84±2.7%

4.3. Delay performance

The comparison results of latency performance are shown in Table 4. In both conventional traffic and adversarial sample injection scenarios, the processing time of HEPS is significantly lower than other models, and the time fluctuation is minimal, indicating that it has more stable real-time response capability while ensuring efficient processing speed.

Table 4. ADT comparison

Scenario	HEPS	ADFA	DualFlow	TAPAS
Regular Traffic	38.12±2.1ms	121.53±8.2ms	79.35±5.1ms	217.53±15.3ms
Adversarial Sample Injection	51.27±3.8ms	143.62±9.8ms	106.53±7.3ms	326.18±21.7ms

5. Conclusion

The HEPS proposed in this article significantly outperforms existing baseline models in terms of accuracy, adversarial robustness, and efficiency in detecting covert attacks through mechanisms such as dual stream processing channels and dynamic weight allocation. This framework effectively breaks through the technical bottleneck of traditional detection methods, provides a reliable solution for target detection in complex attack environments, and also offers new ideas for the deep application of large models in the field of security.

References

[1] Tian, J., Tan, R., Guan, X., & Liu, T. (2018). Enhanced hidden moving target defense in smart grids. IEEE transactions on smart grid, 10(2), 2208-23.
[2] Zhang, J., & Li, C. (2019). Adversarial examples: Opportunities and challenges. IEEE transactions on neural networks and learning systems, 31(7), 2578-93.
[3] Batur Şahin, C., & Abualigah, L. (2021). A novel deep learning-based feature selection model for improving the static analysis of vulnerability detection. Neural Computing and Applications, 33(20), 14049-67.
[4] Yeung, D. Y., & Ding, Y. (2003). Host-based intrusion detection using dynamic and static behavioral models. Pattern recognition, 36(1), 229-43.
[5] Chen, Y., Sun, Z., Zhang, R., Yao, L., & Wu, G. (2023). Attention mechanism based neural networks for structural post-earthquake damage state prediction and rapid fragility analysis. Computers & Structures, 281, 107038.
[6] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023, November). Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In Proceedings of the 16th ACM workshop on artificial intelligence and security (pp. 79-90).