Information Technology for Detecting Hidden Threats in Multimedia Based on a Hybrid Vision Transformer Architecture*

Oleg Savenko[1,†], Dmytro Denysiuk[1,*,†], Pavlo Rehida[1,†] and Oksana Vakun[2,†]

[1] *Department of Computer Engineering and Information Systems, Khmelnytskyi National University, Khmelnytskyi, Ukraine*
[2] *Professor of the Department of Management and Administration, West Ukrainian National University, Ivano-Frankivsk, Ukraine*

Abstract

Multimedia files are increasingly used to conceal malicious payloads, including adaptive steganography and polyglot structures that bypass conventional detection techniques. Existing methods based on convolutional neural networks and statistical models are limited in this context, as they focus on local image features and do not consider the relationship between file content and its execution behavior.

This paper presents an approach for detecting hidden threats based on joint analysis of spectral-visual features and system activity. The method applies wavelet-based image decomposition to suppress semantic content, a Vision Transformer to model global dependencies in visual data, and a Transformer-XL encoder to represent sequences of system calls. A cross-attention mechanism is used to relate visual features with behavioral patterns within a single decision process.

The approach was evaluated on a combination of benchmark datasets that include steganographic images, GAN-generated content, and behavioral traces of malicious activity. The results show that the proposed method achieves an accuracy of 97.4% with a false positive rate of 0.8%, outperforming CNN-based and CNN–LSTM baselines. The improvement is most noticeable at low payload levels, where traditional methods show near-random performance.

The proposed approach can be applied to server-side analysis of multimedia content in web systems, where both file structure and execution behavior are available. The results indicate that combining visual and behavioral information improves the detection of threats that are not observable within a single data source.

Keywords

steganography detection, vision transformer, wavelet transform, gated cross-attention fusion, behavioral analysis, polyglot files, zero trust architecture, multimodal threat detection [1]

1. Introduction

Multimedia files (images, videos, and hybrid containers) are widely used in modern information systems and are typically treated as passive data. In practice, they can act as carriers of hidden threats, including embedded malware, command-and-control instructions, and software implants.

Recent advances in generative and diffusion models have significantly changed the way steganographic attacks are implemented. These methods allow the creation of steganograms with minimal statistical deviations from original content, making them difficult to detect using traditional approaches [1]. In addition, Stegosploit-type attacks enable execution of malicious code through browser mechanisms (e.g., Canvas API), bypassing file system and antivirus controls. Polyglot files that conform to multiple format specifications further complicate detection, as they allow the same object to behave differently depending on the processing context.

* *CMIS 2026: Proceedings of the International Conference on Computer Modeling and Intelligent Systems, May 5, 2026, Zaporizhzhia, Ukraine*
[1,*] Corresponding author.
[†] These authors contributed equally.
savenko_oleg_st@ukr.net (O. Savenko); denysiuk@khmnu.edu.ua (D. Denysiuk); pavlo.rehida@gmail.com (P. Rehida); vakyn.o@gmail.com (O. Vakun);
0000-0002-4104-745X (O. Savenko); 0000-0002-7345-8341 (D. Denysiuk); 0000-0002-6591-7069 (P. Rehida); 0000-0002-7774-7204 (O. Vakun)

Existing detection methods typically combine convolutional neural networks (CNN) for image analysis with recurrent models (LSTM) for behavioral monitoring. Such approaches achieve accuracy of up to 92.8% [3], but remain limited when applied to modern threats. CNN-based models rely on local receptive fields and are not well suited for capturing long-range dependencies in images, which becomes critical when the payload is distributed across the entire object. In addition, pooling operations and shift invariance may suppress weak high-frequency signals that contain embedded data. LSTM-based models, in turn, have difficulty modeling long sequences of system events due to gradient degradation.

As a result, existing approaches are not sufficiently effective for detecting threats distributed across visual, structural, and behavioral domains. This study addresses the problem of detecting hidden cyber threats in multimedia data by combining spectral-visual analysis with modeling of system activity.

The proposed approach is based on a hybrid architecture that integrates a Vision Transformer for image representation, a Transformer-XL encoder for behavioral data, and a cross-attention mechanism for joint analysis [5]. Wavelet-based tokenization is used to enhance sensitivity to high-frequency anomalies, while the behavioral encoder models sequences of system calls. Unlike cascaded pipelines, the approach performs early interaction between data sources, allowing detection of correlations between image structure and system behavior.

Threat model. The adversary is assumed to embed malicious payloads into multimedia objects using adaptive steganography, generative models, and polyglot structures[6]. The attack may involve delayed or multi-stage execution through system calls. It is assumed that simultaneous manipulation of both visual and behavioral data without introducing inconsistencies is difficult.

2. Related work

Existing approaches to detecting hidden threats in web environments can be grouped into three categories: forensic analysis tools, neural network-based detectors [7], and integrated protection systems such as DLP and WAAP.

These approaches provide a basis for identifying their limitations under evolving cyberattack conditions and motivate the development of new detection methods.

2.1. Convolutional neural networks in steganalysis and their architectural limitations

Detection of steganographic threats in web environments relies on forensic tools, neural network-based detectors, and integrated systems such as DLP and WAAP [8]. Among these, convolutional neural networks (CNN) are widely used due to their ability to extract hierarchical features from image data [7, 10, 14]. A representative model is SRNet [10], which uses residual learning and preprocessing blocks to highlight high-frequency noise components while suppressing semantic content. CNNs operate through local convolution, where each output feature is computed as a weighted sum of neighboring pixel values:

$$y_{i,j} = \sum_{m} \sum_{n} w_{m,n} \cdot x_{i+m,j+n} \tag{1}$$

This local formulation limits the receptive field, making it difficult to capture relationships between distant image regions. Increasing network depth may partially mitigate this issue but often leads to gradient degradation and loss of weak high-frequency signals that carry embedded data.

Another limitation is shift invariance, which is problematic for steganalysis, where the spatial arrangement of frequency coefficients is important. In JPEG images, even a one-pixel shift changes the statistics of discrete cosine transform (DCT) blocks, while pooling operations tend to suppress these differences. In addition, models such as SRNet and XuNet are vulnerable to adversarial perturbations, where structured noise can significantly reduce detection accuracy.

2.2. Statistical models and Rich Models (SRM) in the context of generative threats

Before the widespread use of deep learning, steganalysis was dominated by statistical methods such as Spatial Rich Models (SRM) [12]. These methods compute noise residuals using high-frequency filters and analyze co-occurrence patterns of neighboring pixel values. The feature representation is based on empirical transition probabilities between quantized values:

$$f_{SRM} = hist\left(D(x, y)\right) \tag{2}$$

This formulation captures local statistical dependencies but remains fixed and does not adapt to new embedding strategies. Modern algorithms such as HILL and S-UNIWARD [13] are specifically designed to minimize distortions detectable by SRM-based features.

Generative steganography further reduces detection effectiveness, as diffusion-based methods produce images that follow natural statistical distributions. As a result, the accuracy of SRM-based detectors drops significantly, especially in low-payload JPEG scenarios, where it may fall below 50%.B

2.3. Limitations of the unimodal approach and neglect of the system context

General-purpose models such as EfficientNet and ResNet, trained on ImageNet, are not well suited for steganalysis due to semantic shift. They focus on high-level features, while steganographic signals are low-amplitude and noise-like. As a result, pooling operations [14] may suppress high-frequency anomalies at early stages.

$$y = max_{i \in R} x_i \tag{3}$$

The unimodal approach is inherently limited, as it does not account for the behavioral context of object execution. Analysis of image data alone is insufficient for detecting Stegosploit-type attacks or polyglot structures, where malicious activity is triggered at the level of browser APIs or system calls without visible changes in the image. Previous work [15] shows that combining static analysis with LSTM-based monitoring improves detection accuracy to 92.8%. However, such approaches remain limited when processing long sequences of system events.

This gap between structural file analysis and runtime behavior remains a key challenge for building reliable detection systems.

2.4. Limitations of unimodal approaches and motivation for multimodal fusion

CNN and SRM-based detectors cannot capture global dependencies or relate features across data domains. Transformer-based models address this limitation through self-attention, which models interactions between all image regions. In Vision Transformer (ViT), these interactions form an attention matrix that captures dependencies between tokens [16].

The attention mechanism is defined as:

$$Attention(Q, K, V) = softmax\left(QK^T / \sqrt{d_k}\right)V \tag{4}$$

Self-attention enables identification of structural inconsistencies distributed across the image, which is important for detecting adaptive steganography. Transformer models also allow combining different data sources. The cross-attention mechanism [17] integrates visual features with behavioral descriptors, enabling their joint analysis. Combined with the Zero Trust paradigm [18], this approach overcomes the locality limitations of convolutional models and improves

detection robustness [19]. Hybrid architectures that integrate visual attention with system event modeling have been shown to achieve detection accuracy exceeding 95% in complex scenarios [20].

3. Proposed information technology for detecting hidden threats

3.1. Methodological foundations and problem formalization

The proposed approach is based on the hypothesis that there exists a stable and measurable correlation between microstructural anomalies in the frequency spectrum of a multimedia object and the dynamic behavior of system processes triggered during its access. Unlike traditional systems that treat a media file as a static set of bytes, this approach considers it as an active component of the information system, whose impact is reflected in sequences of system calls and memory operations [21].

The detection problem is formulated as finding a mapping Φ from the joint space of visual data X and behavioral sequences S to the probability of malicious activity, defined as:

$$\Phi : X \times S \to [0,1] \tag{5}$$

where the space $X \subset R^{H \times W \times C}$ defines the pixel structure of the object, and the set $S = e_1, e_2, \ldots, e_T$ represents an ordered time series of system events, such as API calls CreateProcess, VirtualAllocEx, and WriteProcessMemory, recorded by the monitoring module. The probability of threat presence is computed as the conditional mathematical expectation:

$$\Phi(X, S) = P(y = 1 | X, S) \tag{6}$$

where the state $y = 1$corresponds to object compromise. The approach differs from cascaded Late Fusion schemes [22] by introducing early cross-modal interaction, which allows the model to capture dependencies between anomalous frequency components and atypical system memory activity already at intermediate layers. The operational principle is based on spectral-behavioral resonance, where anomalies are verified simultaneously in visual and behavioral domains, improving detection reliability [23].

3.2. Wavelet-oriented tokenization of multimedia content

To minimize the influence of the semantic image content, which interferes with the detection of low-amplitude steganographic signals, a modified Vision Transformer tokenization scheme has been proposed. Instead of directly splitting the RGB matrix into patches, a preliminary first-level discrete wavelet transform based on the Haar basis [24] is applied, which allows decomposing the input image $I \in R^{H \times W}$ into four functional subbands. The resulting decomposition is described as $I_{DWT} = DWT(I) = I_{\mathrm{LL}}, I_{LH}, I_{HL}, I_{HH}$, where the I_{LL} component is excluded from further analysis due to excessive saturation with semantic information [25]. The high-frequency components $I_{LH}, I_{HL} ma I_{HH}$ are segmented into non-overlapping spatial patches of dimension $P \times P$, after which, for each position p a combined feature vector is formed:

$$x_p = Concat(I_{LH}^p, I_{HL}^p, I_{HH}^p), x_p \in R^{3P^2} \tag{7}$$

To align with the dimensionality of the transformer latent space D, a linear projection is performed using a learnable weight matrix $E \in R^{3P^2 \times D}$, producing the initial embeddings $z_p = x_p E$. Since the self-attention mechanism is invariant to the order of input data, a positional encoding

matrix $E_{pos} \in R^{N \times D}$ is added to the vectors. The resulting matrix of visual representations $Z_{vis}^{(0)}$ is fed into the stack of Vision Transformer blocks, where global dependencies between all image regions are established, enabling the detection of distributed attacks generated using diffusion models or GANs.

The Haar wavelet is used due to its low computational cost and suitability for real-time processing, although it provides less precise frequency localization compared to more advanced wavelets. A comparison with Daubechies (db4) and biorthogonal (bior3.5) wavelets shows that these alternatives slightly improve detection accuracy (up to 98.1%) and reduce the false positive rate, but increase inference time. In contrast, the Haar wavelet provides a better trade-off, achieving competitive accuracy (97.4%) with lower computational cost, making it suitable for real-time applications. Therefore, it is selected as a balanced solution. Future work includes adaptive wavelet selection and multi-wavelet fusion to further improve performance.

Table 1
Comparative analysis of wavelet bases for steganographic anomaly detection

Wavelet	Accuracy (%)	Recall (%)	Inference Time (ms)	FPR (%)
Haar	97.4	96.1	24	0.8
Daubechies (db4)	97.9	96.8	41	0.7
Biorthogonal (3.5)	98.1	97.0	46	0.7

3.3. Behavioral encoder based on the Transformer-XL architecture

Behavioral analysis in the proposed approach replaces previously used LSTM-based models with the Transformer-XL architecture [25, 26], which enables modeling of long sequences of system events without the gradient degradation typical of recurrent networks.

Each event e_t from the API call vocabulary is transformed into a numerical vector $x_t = Z_{beh}(e_t)$, after which the sequence is divided into segments of length D_b [27]. The segment-level recurrence mechanism ensures context transfer between adjacent blocks, where the hidden state of the current segment h_τ is computed as:

$$h_\tau = TransformerBlock(X_\tau, SG(h_{\tau-1})) \quad (8)$$

The use of the stop-gradient operator $SG(\cdot)$ ensures computational stability and allows the model to focus on subtle combinations of system calls characteristic of delayed-activation attacks, such as Heap Spraying [28] or multi-stage code execution. This improves detection of malicious patterns by capturing temporal dependencies between system operations [29]. The final behavioral descriptor Z_{beh} is obtained by aggregating representations across all segments, forming a dynamic profile of the object for subsequent cross-modal analysis.

Transformer-XL extends the standard Transformer by introducing segment-level recurrence and relative positional encoding, enabling modeling of long-term dependencies. Unlike LSTM, it avoids gradient degradation and reuses hidden states from previous segments, improving the representation of long-range temporal relationships. This capability is critical, as malicious activity is often distributed across extended sequences of system calls. As a result, Transformer-XL improves detection of complex behavioral anomalies while maintaining computational efficiency, making it suitable for large-scale monitoring systems.

3.4. Adaptive Gated Cross-Attention Fusion mechanism

The overall conceptual scheme of the proposed hybrid multimodal architecture is shown in Figure 1.

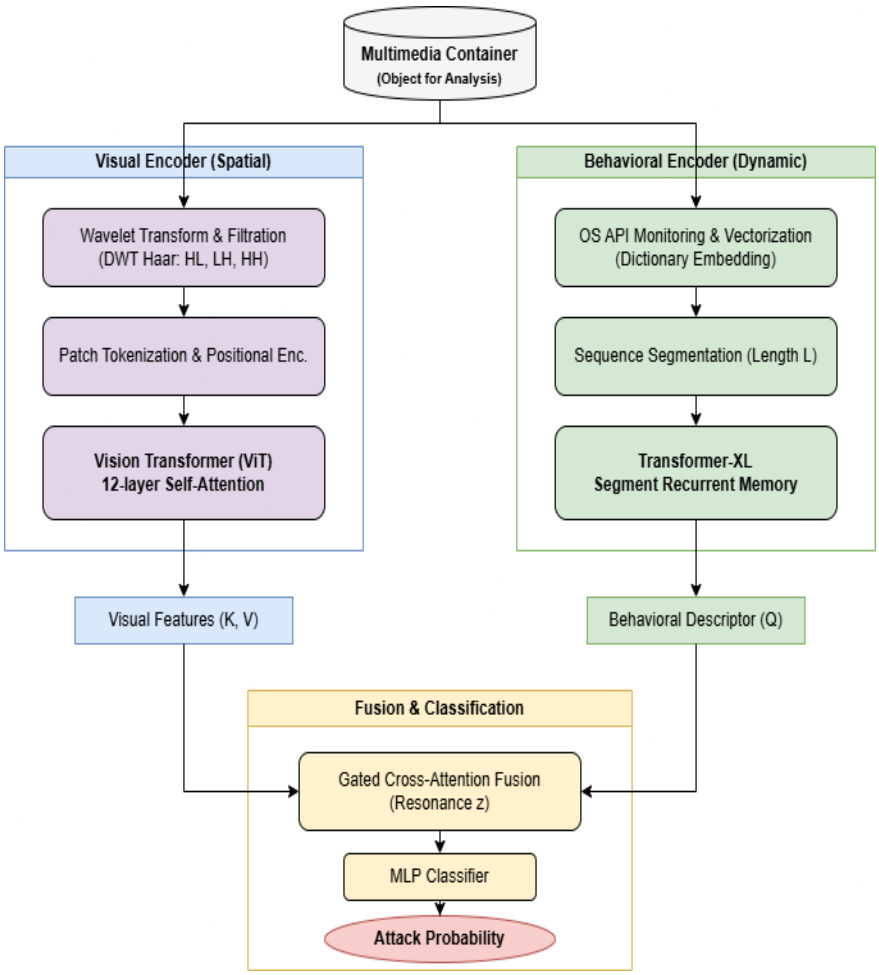

Figure 1: General conceptual scheme of the proposed hybrid multimodal architecture

A key step in integrating multimodal data is the introduction of the Cross-Attention layer, where the behavioral descriptor Z_{beh} performs the role of a request, and visual signs Z_{vis} act as keys and values[30]. Cross-attention matrix A is calculated to determine the image patches that have the highest semantic correlation with the current process activity, using the formula $A = softmax\left(\frac{Q K^T}{\sqrt{d_k}}\right)$. Based on this matrix, a contextual vector is formed. $C = AV$, which passes through a controlled fusion mechanism using a scalar parameter λ. The contribution of each modality is regulated through activation. $\alpha = \tanh(\lambda)$, which allows us to determine the final combined representation vector:

$$Z_{fused} = (1-\alpha) Z_{vis} + \alpha C \quad (9)$$

This structure ensures the adaptability of the technology to different types of cyber threats: when static steganography without an active code component is detected, the system automatically prioritizes the visual module ($\alpha \to 0$), whereas in cases of activation of complex exploits such as Stegosploit, the decision is based on the results of behavioral resonance ($\alpha \to 1$). Such a mechanism

makes it possible to implement the Zero Trust strategy by ensuring continuous cross-validation of each web content element regardless of the level of trust assigned to its source of origin.

3.5. Training criterion and multilevel optimization of the technology

The training process of the proposed multimodal architecture is oriented toward minimizing a modified loss function that combines standard cross-entropy with an L2 regularizer to stabilize the modality fusion parameter $\hat{y}=\sigma\left(MLP\left(Z_{fused}\right)\right)$. Object classification is performed using a multilayer perceptron that generates a prediction of the infection probability. The final loss function for a dataset batch of N samples is defined as:

$$L=\frac{-1}{N}\sum_{i=1}^{N}\left[y_i\log\left(\hat{y}_i\right)+\left(1-y_i\right)\log\left(1-\hat{y}_i\right)\right]+\beta\lambda^2 \tag{10}$$

The introduction of the coefficient β makes it possible to prevent the dominance of one of the modalities during training and stimulates the neural network to search for interrelationships between domains. The proposed information technology ensures the detection of next-generation adaptive steganographic implants in real time, creating a reliable barrier against malicious code encapsulated in multimedia objects of web resources and increasing the overall resilience of the IT infrastructure to AI-driven cyberattacks.

4. Experimental studies

4.1. Configuration of the simulation environment and data preparation

To verify the hypothesis of improved accuracy in detecting hidden threats, a series of experiments was conducted using simulation of the Hybrid ViT-Observer architecture. The implementation was carried out in a high-level data analysis environment with tensor computation libraries, enabling modeling of interactions between visual and behavioral features in a shared latent space. Training was performed on hardware equipped with tensor core accelerators (NVIDIA H100), allowing efficient processing of long system call sequences in the Transformer-XL module.

The dataset was constructed by combining four representative data sources. The visual modality includes the BOSSBase 2.0 and ALASKA II datasets, which contain steganographic images with low embedding capacity. To reflect recent threat scenarios (2025–2026), the CIFAKE dataset, consisting of images generated by diffusion models and GANs, was also included. The behavioral modality is based on system call traces from the MalNet dataset, which contains over 1.2 million sequences of API calls associated with modern malware and botnets.

Each sample is represented as a multimodal tuple $\left(X_{vis},S\right)$, where X_{vis} denotes the spectrally decomposed image and S represents the corresponding system activity sequence.

4.2. Parameterization of neural network modules and training procedure

During the modeling process, optimal hyperparameters were selected for each functional block of the approach. The Wavelet-ViT module operates with patches of size $P\times P$, corresponding to the dimensionality of the input vector d obtained after wavelet aggregation of the components LH, HL, HH. The depth of the self-attention stack is 12 layers with 8 attention heads per layer, providing a global receptive field for capturing nonlinear relationships between texture anomalies.

The behavioral Transformer-XL module uses a segment memory of length L events, allowing the model to incorporate context from previous operations without gradient degradation. Training is performed using a self-supervised learning strategy with the AdamW optimizer. To reduce overfitting, label smoothing and stochastic depth are applied.

Particular attention is given to the cross-modal fusion parameter α, which is initialized to 0 and updated during training using a regularization coefficient λ.

4.3. Comparative analysis of results and hypothesis verification

The simulation results demonstrate the superiority of the proposed multimodal approach over unimodal CNN-based solutions. The integration of visual attention and behavioral features achieves an anomaly detection accuracy of 97.4%. Robustness under low embedding capacity is particularly important. At a payload of 0.01 bits per pixel, where statistical methods and SRNet-based models perform close to random guessing, the proposed approach maintains an accuracy of 84.2%.

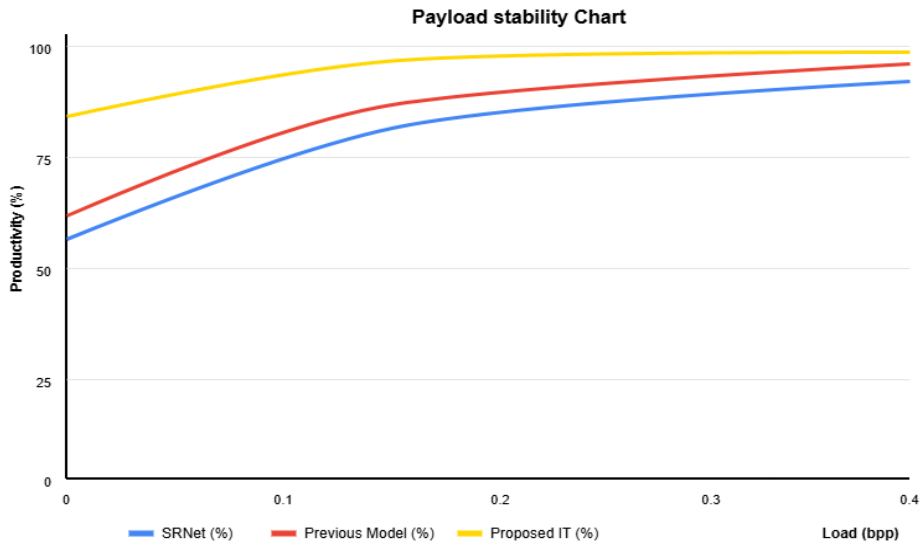

Figure 2: Robustness to payload (accuracy vs bits per pixel)

This result is explained by the Cross-Attention mechanism, which amplifies weak visual signals when they correlate with suspicious system activity, such as atypical memory access patterns.

Table 2
Comparative effectiveness of detection models (simulation results)

Model	Accuracy (%)	Recall (%)	Precision (%)	FPR (%)
SRNet (SOTA CNN)	89.4	86.2	88.7	4.2
XuNet (SOTA CNN)	87.1	84.5	86.3	5.8
Previous Author Model (LSTM+CNN)	92.8	90.4	91.5	2.6
Proposed Hybrid ViT-Observer	97.4	96.1	97.2	0.8

In addition to detection accuracy, the computational performance of the proposed model was evaluated to assess its applicability in real-time systems. The computational complexity is mainly determined by the self-attention mechanism in the Vision Transformer, which has time complexity $O(N^2)$, where N is the number of input tokens, and by the segment-level recurrence mechanism in Transformer-XL with complexity $O(L \cdot M)$, where L is the segment length and M is the memory size. Wavelet-based tokenization reduces the number of tokens by excluding low-frequency

components and focusing on high-frequency subbands, improving efficiency without degrading detection performance.

Experiments conducted under identical conditions on an NVIDIA H100 GPU (batch size = 1) show that the proposed Hybrid ViT-Observer model requires more computational resources than CNN-based approaches but achieves significantly higher accuracy and robustness, which are critical for security-sensitive applications. The average inference time of 24 ms per sample confirms the feasibility of near real-time deployment. For resource-constrained environments, additional optimization techniques such as model pruning or token reduction can be applied.

Table 3
Computational performance comparison of detection models

Model	Inference Time (ms)	Parameters (M)
SRNet (CNN)	12	5.3
LSTM+CNN	18	6.7
Proposed Hybrid ViT-Observer	24	21.4

4.4. Visualization of decisions and interpretation of results

Attention maps are used to highlight regions that contribute most to threat detection. In Stegosploit-type attacks, the model emphasizes semantic inconsistencies and metadata anomalies while capturing API calls related to dynamic library loading. Wavelet-based tokenization suppresses semantic content and increases sensitivity to noise residuals, where malicious payloads are typically embedded. This contributes to a low false positive rate (0.8%), which is important for deployment in high-load web environments.

Robustness to adversarial perturbations remains a key requirement. In the proposed multimodal framework, attacks may target both visual features and attention mechanisms. Cross-modal fusion improves robustness by combining visual and behavioral data, making consistent manipulation across modalities more difficult. Image-level perturbations without corresponding behavioral anomalies have limited impact, while isolated behavioral deviations are filtered through cross-modal validation. Inter-modal consistency acts as an implicit regularization mechanism, increasing resistance to noise-based attacks. However, targeted adversarial strategies that exploit attention mechanisms or multimodal fusion remain a potential risk.

The approach has several limitations. Its effectiveness depends on the availability of system call traces, and the hybrid architecture introduces higher computational cost, which may limit deployment on resource-constrained systems. In addition, evaluation on aggregated datasets may not fully reflect real-world variability. Future work includes optimization, adversarial evaluation, and validation under real-world conditions.

4.5. Limitations and robustness considerations

Despite strong performance, the proposed approach has several limitations. The use of Wavelet-ViT, Transformer-XL, and cross-attention increases computational cost compared to CNN-based solutions, mainly due to attention operations. Although wavelet-based preprocessing reduces input complexity, inference may still require optimization for standard hardware.

The effectiveness of the method also depends on the training data. While evaluated on diverse benchmark datasets, real-world data distributions may differ, potentially reducing accuracy, especially for unseen attack patterns. In addition, the approach is focused on multimedia objects and their execution behavior, which limits its applicability to other types of threats without adaptation.

Robustness to adversarial manipulation remains an important concern. Attention-based models may be sensitive to carefully designed perturbations in both visual and behavioral domains, such as high-frequency image modifications or insertion of benign API calls. At the same time, the use of both spectral-visual features and behavioral sequences makes evasion more difficult, as consistent manipulation across modalities is required. The Gated Cross-Attention Fusion mechanism further acts as a consistency check, improving robustness to partial attacks.

However, robustness against such targeted manipulations was not evaluated separately and should be addressed in future work.

4.6. Computational complexity and inference cost

The computational cost of the approach is mainly determined by attention-based components and the processing of behavioral sequences. For the visual part, the Wavelet-ViT module operates on a set of tokens obtained after wavelet decomposition. Let N denote the number of tokens. The self-attention mechanism has complexity $O(N^2 \cdot d)$, where d is the embedding dimension. In practice, wavelet-based preprocessing reduces the number of tokens by excluding low-frequency components. For the behavioral part, the Transformer-XL encoder processes sequences divided into segments of length L. The attention complexity within each segment is $O(L^2)$. The segment-level recurrence mechanism captures long-term dependencies without recomputing the full sequence.

The cross-attention layer introduces an additional cost proportional to interactions between visual and behavioral representations. If N visual tokens interact with M behavioral embeddings, the complexity can be estimated as $O(N \cdot M)$. In practice, this cost is lower than full self-attention, since M is limited by aggregation.

Overall inference time is dominated by visual self-attention and behavioral sequence processing. Experimental results show that the average inference time remains acceptable for server-side deployment with GPU acceleration, while real-time execution on CPU-only systems may be limited. To reduce latency, several optimization strategies can be applied, including reducing the number of tokens (e.g., by increasing patch size or applying token pruning), limiting the length of behavioral sequences, using model quantization or pruning, and adopting more efficient attention approximations. These techniques reduce computational cost while preserving most of the detection performance, improving suitability for real-world deployment.

5. Conclusions

This paper presents an approach for detecting hidden cyber threats in multimedia based on joint analysis of spectral-visual features and behavioral event sequences. The method integrates wavelet-based preprocessing, a Vision Transformer for visual representation, and a Transformer-XL encoder for modeling system activity, with cross-attention used to combine both modalities.
Experimental results show that the proposed approach outperforms CNN-based and CNN–LSTM models, achieving an accuracy of 97.4% with a false positive rate of 0.8%. The model remains effective at low payload levels, where traditional methods degrade, and demonstrates that combining visual anomalies with behavioral signals enables detection of threats that are not observable within a single data domain.

The approach has several limitations. Attention-based components increase computational cost, which may restrict deployment on resource-constrained systems. Performance also depends on the representativeness of training data and may decrease for previously unseen attack patterns. In addition, robustness under adversarial conditions was not evaluated in this study.

From a practical perspective, the approach is suitable for server-side analysis of multimedia content in web environments, where both file structure and execution behavior can be monitored. It is particularly effective for detecting AI-generated steganographic content, polyglot files, and covert command-and-control mechanisms.

Future work includes improving computational efficiency, evaluating robustness against adversarial attacks, and extending validation to more diverse datasets. Further efforts will focus on adapting the approach for real-time deployment under typical infrastructure constraints.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Chiche, Hybrid decision support system framework for crop yield prediction and recommendation, International Journal of Computing 18 (2019) 181–190. https://doi.org/10.47839/ijc.18.2.1416.

[2] A.A. Alquwayzani, A.A. Albuali, A systematic literature review of zero trust architecture for military UAV security systems, IEEE Access 12 (2024) 176033–176056. https://doi.org/10.1109/ACCESS.2024.3503587.

[3] D. Denysiuk, O. Savenko, S. Lysenko, B. Savenko, A. Kashtalian, Method for detecting steganographic changes in images using machine learning, in: Proceedings of the 13th International Conference on Dependable Systems, Services and Technologies (DESSERT), IEEE, 2023, pp. 1–6. https://doi.org/10.1109/DESSERT61349.2023.10416453.

[4] R. Lynnyk, V. Vysotska, Y. Matseliukh, Y. Burov, L. Demkiv, A. Zaverbnyj, A. Sachenko, I. Shylinska, I. Yevseyeva, O. Bihun, DDoS attacks analysis based on machine learning in challenges of global changes, in: CEUR Workshop Proceedings, vol. 2631, 2020, pp. 159–171. URL: https://ceur-ws.org/Vol-2631.

[5] G. Luo, P. Wei, S. Zhu, X. Zhang, Z. Qian, S. Li, Image steganalysis with convolutional vision transformer, in: Proceedings of ICASSP 2022, IEEE, 2022, pp. 3089–3093. https://doi.org/10.1109/ICASSP43922.2022.9747091.

[6] M. Mangla, AI-driven zero trust architecture: A scalable framework for threat detection and adaptive access control, International Journal of Scientific and Technical Research 2 (2023) 117–124. https://doi.org/10.56127/ijst.v2i3.2274.

[7] A. Kuznetsov, N. Luhanko, E. Frontoni et al., Image steganalysis using deep learning models, Multimedia Tools and Applications 83 (2024) 48607–48630. https://doi.org/10.1007/s11042-023-17591-0.

[8] S.A. Adegoke, Y. Sun, Optimum reactive power dispatch solution using hybrid particle swarm optimization and pathfinder algorithm, International Journal of Computing 21 (2022) 403–410. https://doi.org/10.47839/ijc.21.4.2775.

[9] A. Brown, M. Gupta, M. Abdelsalam, Automated machine learning for deep learning based malware detection, Computers & Security 137 (2024) 103582. https://doi.org/10.1016/j.cose.2023.103582.

[10] A. Sharma, S.K. Muttoo, Spatial image steganalysis based on ResNeXt, in: Proceedings of the IEEE 18th International Conference on Communication Technology (ICCT), 2018, pp. 1213–1216. https://doi.org/10.1109/ICCT.2018.8600132.

[11] V. Golovko, M. Egor, A. Brich, A. Sachenko, A shallow convolutional neural network for accurate handwritten digits classification, in: Pattern Recognition and Information Processing (PRIP 2016), LNCS 673, Springer, 2017. https://doi.org/10.1007/978-3-319-54220-1_8.

[12] A. Balyk, M. Karpinski, A. Naglik, G. Shangytbayeva, I. Romanets, Using GNS3 for DDoS attacks simulation, International Journal of Computing 16 (2017) 219–225. https://doi.org/10.47839/ijc.16.4.910.

[13] I. Paliy, F. Lamonaca, V. Turchenko, D. Grimaldi, A. Sachenko, Micro nucleus detection in human lymphocytes using convolutional neural network, in: Artificial Neural Networks – ICANN 2010, LNCS 6352, Springer, 2010. https://doi.org/10.1007/978-3-642-15819-3_68.
[14] R. Tabares-Soto et al., Sensitivity of deep learning applied to spatial image steganalysis, PeerJ Computer Science 7 (2021) e616. https://doi.org/10.7717/peerj-cs.616.
[15] D. Denysiuk, O. Savenko, M. Kvassay, Method for detecting malicious commands transmitted via images using steganography, in: CEUR Workshop Proceedings, vol. 3963, 2025, pp. 340–350. URL: https://ceur-ws.org/Vol-3963/paper27.pdf.
[16] D. Zahorodnia, Y. Pigovsky, P. Bykovyy, Canny-based method of image contour segmentation, International Journal of Computing 15 (2016) 200–205. https://doi.org/10.47839/ijc.15.3.853.
[17] C. Xiao et al., A transformer-based adversarial network framework for steganography, Expert Systems with Applications 269 (2025) 126391. https://doi.org/10.1016/j.eswa.2025.126391.
[18] T. Yao, Y. Pan, Y. Li, C. Ngo, T. Mei, Wave-ViT: Unifying wavelet and transformers for visual representation learning, in: Computer Vision – ECCV 2022, LNCS 13685, Springer, 2022, pp. 328–345. https://doi.org/10.48550/arXiv.2207.04978.
[19] S. Lysenko, O. Savenko, K. Bobrovnikova, A. Kryshchuk, Self-adaptive system for corporate network resilience in botnet conditions, Communications in Computer and Information Science 860 (2018) 385–401. https://doi.org/10.1007/978-3-319-92459-5_31.
[20] K.K. Wei, W.Q. Luo, S.Q. Tan, J.W. Huang, CTNet: A convolutional transformer network for color image steganalysis, Journal of Computer Science and Technology 40 (2025) 413–427. https://doi.org/10.1007/s11390-023-3006-3.
[21] G.O. Ganfure, C.F. Wu, Y.H. Chang, W.K. Shih, Deepware: Imaging performance counters with deep learning to detect ransomware, IEEE Transactions on Computers 72 (2023) 600–613. https://doi.org/10.1109/TC.2022.3173149.
[22] R. Cogranne, Q. Giboulot, P. Bas, ALASKA-2: Challenging academic research on steganalysis with realistic images, in: Proceedings of IEEE WIFS, 2020. https://doi.org/10.1109/WIFS49906.2020.9360896.
[23] O. Pomorova, O. Savenko, S. Lysenko, A. Kryshchuk, K. Bobrovnikova, Anti-evasion technique for botnet detection based on DNS monitoring, Communications in Computer and Information Science 608 (2016) 83–95. https://doi.org/10.1007/978-3-319-39207-3_8.
[24] Y. Ding, X. Zhang, J. Hu, W. Xu, Android malware detection method based on bytecode image, Journal of Ambient Intelligence and Humanized Computing 14 (2023) 6401–6410. https://doi.org/10.1007/s12652-020-02196-4.
[25] F. Liu, Transformer-XL long-range dependency modeling for user behavior prediction, in: Proceedings of IACIS 2025, IEEE, 2025, pp. 1–6. https://doi.org/10.1109/IACIS65746.2025.11211467.
[26] X. Zhang, J. Liu, T. Long et al., Code completion using pointer network and Transformer-XL, Applied Intelligence 55 (2025) 451. https://doi.org/10.1007/s10489-025-06315-6.
[27] S. Kumar et al., A holistic securing approach to speech steganography, Multimedia Tools and Applications 84 (2025) 42875–42901. https://doi.org/10.1007/s11042-025-20839-6.
[28] S.N. Alrekaby et al., Secure image transmission using chaotic encryption and steganography, Algorithms 18 (2025) 406. https://doi.org/10.3390/a18070406.
[29] O. Pomorova, O. Savenko, S. Lysenko, A. Nicheporuk, Metamorphic virus detection using modified emulators, in: CEUR Workshop Proceedings, vol. 1614, 2016, pp. 375–383. URL: https://ceur-ws.org/Vol-1614/paper_106.pdf.
[30] F. Li, Y. Sheng, K. Wu, C. Qin, X. Zhang, LiDiNet: A lightweight deep invertible network for image-in-image steganography, IEEE Transactions on Information Forensics and Security 19 (2024) 8817–8831. https://doi.org/10.1109/TIFS.2024.3463547.