e-ISSN: 3079-6377 p-ISSN: 3079-6369

Frequency-Domain Characterization of Memory Poisoning Propagation in Multi-Agent Collaborative Systems

Wei Ming Tan[1], Cheryl Lim[2], Jonathan Lee[3]*

Department of Computer Science, National University of Singapore, 13 Computing Drive, Singapore 117417

Corresponding author: jonathan.lee@placeholder.edu

Abstract

Memory poisoning in multi-agent collaborative systems may exhibit structured propagation patterns across communication channels. This study analyzes poisoning diffusion using frequency-domain decomposition of inter-agent memory update signals. Memory state transitions are transformed via discrete Fourier analysis to identify anomalous high-frequency components introduced by adversarial perturbations. A spectral attenuation filter selectively suppresses abnormal frequency bands before synchronization. Experiments were conducted on 120-agent distributed coordination tasks over 10,000 update cycles. Poisoning injection at 15% intensity increased high-frequency spectral energy by 63.2% compared to clean baselines. Spectral filtering reduced contamination spread from 58.4% to 19.6% of agents and restored task performance to 87.3% of nominal levels. Frequency-domain modeling provides a novel perspective on poisoning detection and containment in collaborative agent systems.

Keywords

Multi-agent systems; memory poisoning; spectral analysis; frequency-domain filtering; collaborative robustness; signal decomposition

1.Introduction

Multi-agent collaborative systems are increasingly used in distributed coordination, shared planning, autonomous control, and emerging agent-based AI workflows. In these environments, agents depend not only on local observations, but also on repeated memory exchange to maintain state updates, coordination signals, and intermediate decisions across interaction rounds. This shared memory mechanism improves continuity and collective efficiency, yet it also creates a persistent channel through which corrupted information can influence subsequent coordination. Once the memory update process is poisoned, the effect may remain active over multiple synchronization cycles and gradually spread through later interactions. Recent studies on memory poisoning in collaborative agent environments, long-horizon memory attacks in LLM-based agents, and adversarial interference in multi-agent learning have shown that memory is no longer a passive storage component, but a critical attack surface in modern agent systems [1]. In particular, when agents repeatedly retrieve, revise, and synchronize stored information, malicious perturbations may accumulate over time and reshape the collective state of the entire system rather than only damage an isolated memory record [2].

e-ISSN: 3079-6377 p-ISSN: 3079-6369

This risk is especially important in collaborative settings where memory updates are repeatedly transmitted across communication links and incorporated into later decisions. Existing work has shown that corrupted memory may propagate across agents, alter downstream reasoning, and remain hidden until later synchronization or planning stages [3,4]. Compared with one-step adversarial perturbation, memory poisoning in collaborative systems has a stronger temporal carryover effect and a broader structural impact, because contamination can be preserved, reused, and amplified during repeated coordination. As a result, the key challenge is not only whether poisoned content can be detected at one time point, but also how abnormal memory signals evolve and diffuse across an interacting population over many update cycles. This propagation-oriented view remains insufficiently studied in current research on multi-agent robustness [5,6]. Most existing studies on poisoning resilience still focus on defense in the time domain. Representative methods usually identify abnormal updates from raw temporal sequences, prediction residuals, reconstruction gaps, or attention-based inconsistency patterns. In multivariate time-series anomaly detection, recent models include transformer-based, graph-based, and memory-guided architectures that improve the identification of abnormal patterns in complex temporal observations [7,8]. These methods have achieved strong performance in industrial monitoring, sensor streams, cyber-physical systems, and other sequential anomaly detection tasks. Even so, they were largely developed for general temporal irregularity detection rather than for poisoning propagation in shared agent memory. Their modeling targets are commonly pointwise deviations, local sequence inconsistency, or global forecasting error, whereas poisoning diffusion in collaborative agents is inherently relational and accumulative. The abnormality is not only embedded in the temporal change of one signal, but also in the way corrupted updates move through interacting memory channels. Recent work suggests that frequency-domain analysis may offer a useful complementary view for this problem [9,10]. Time-frequency methods can reveal latent oscillation patterns, periodic shifts, and abrupt disturbances that are difficult to isolate in direct sequence modeling. In multivariate anomaly detection, time-frequency contrastive learning, dual-domain masking, and related spectral methods have shown that anomalous patterns may become more distinguishable after decomposition into different frequency bands [11,12]. Related studies have also demonstrated the value of graph-wavelet analysis, spectral graph anomaly detection, and frequency-domain filtering in networked settings where abnormal behavior is distributed across connected structures rather than concentrated in a single channel [13,14]. These findings are highly relevant to multi-agent memory poisoning, because adversarial perturbations often introduce sudden local inconsistency, irregular synchronization energy, and structurally uneven disturbance during diffusion. Such effects may not always be clearly separated in the raw time domain, but may appear more explicitly as abnormal energy concentration in specific spectral bands. A related research direction has examined robustness in multi-agent communication and coordination. Recent surveys indicate that communication quality strongly affects cooperative multi-agent reinforcement learning performance and that adversarial disruption in shared information can rapidly degrade group behavior [15]. Other studies on cooperative attacks and robust communication mechanisms further show that coordinated policies are sensitive to structured disturbance in message passing, shared memory exchange, and collective state aggregation [16,17]. Most of these studies, however, evaluate attack effects through downstream indicators such as reward loss, coordination failure, reduced policy quality, or communication collapse. Comparatively less attention has been given to the signal structure of poisoning itself, particularly whether poisoned memory updates exhibit identifiable spectral signatures while diffusing across agents. Without such characterization, it remains difficult to explain why some poisoned updates propagate widely whereas others remain locally bounded, and it is also difficult to

e-ISSN: 3079-6377 p-ISSN: 3079-6369

design defenses that intervene before contamination becomes embedded in later synchronization rounds. Several limitations in the current literature motivate further investigation. Existing studies on memory poisoning mainly emphasize attack success, defense accuracy, or final task degradation, while the intermediate diffusion process of corrupted memory across a collaborative population remains underexplored. Current anomaly detection research is still dominated by time-domain modeling, although recent reviews indicate that cross-channel dependence, graph-aware abnormality, and time-frequency mismatch remain unresolved issues in multivariate anomaly analysis [18]. In addition, many benchmark evaluations rely on industrial datasets, generic anomaly detection corpora, or single-agent memory attack settings. These experimental conditions do not fully reflect the structured propagation behavior of poisoned memory in large-scale multi-agent collaboration. Another unresolved issue is whether spectral filtering can suppress abnormal diffusion without excessively damaging normal coordination signals. For collaborative systems, this balance is essential, because over-aggressive filtering may weaken the very information exchange needed for stable cooperation. Against this background, the present study investigates memory poisoning propagation from a frequency-domain perspective and explores whether spectral characterization can provide an effective basis for early intervention. The central assumption is that poisoning diffusion leaves structured traces in inter-agent memory update signals, and that these traces can be identified before synchronization through spectral decomposition of memory state transitions. Based on this idea, the proposed framework transforms memory update sequences into frequency components, detects abnormal high-frequency energy associated with adversarial perturbation, and applies a spectral attenuation mechanism to suppress suspicious bands before corrupted content spreads further through collective synchronization. This design shifts the analytical focus from task-level failure observation to the signal structure of poisoning propagation, and places the defense mechanism directly on the memory update channel where contamination enters shared coordination. The study therefore aims to clarify whether frequency-domain features can reveal the diffusion pattern of poisoned memory, whether spectral filtering can effectively reduce the spread of contamination, and whether such intervention can preserve sufficient signal fidelity for later collaborative decision-making. From a broader perspective, this work contributes to the understanding of poisoning resilience in multi-agent systems by connecting adversarial memory security with time-frequency signal analysis, offering a more interpretable and propagation-aware direction for protecting long-horizon collaborative intelligence.

2.Materials and Methods

2.1. Sample and Task Environment

The study was carried out in a distributed coordination simulation designed for multi-agent collaboration with repeated memory exchange. The system included 120 agents. Each agent kept a local memory stream that recorded recent state updates, coordination messages, and task-related transition values used in later synchronization. The experiments covered 10,000 update cycles under the same communication topology, synchronization interval, and task schedule. The study focused on the spread of poisoned memory updates through inter-agent channels during collaborative execution. Poisoning was injected into memory updates at a fixed intensity of 15% to examine how adversarial perturbations changed the spectral structure of update signals. All runs were performed under the same task conditions so that changes in spectral energy, contamination spread, and task performance could be linked to poisoning and filtering rather than to unrelated system variation.

e-ISSN: 3079-6377 p-ISSN: 3079-6369

2.2. Experimental Design and Control Setting

A controlled comparison was used to test the proposed frequency-domain filtering method against an unfiltered baseline. In the control setting, poisoned memory updates were allowed to pass through the synchronization process without spectral suppression. In the experimental setting, memory update signals were transformed into the frequency domain before synchronization, and abnormal frequency bands were attenuated by a spectral filter before the filtered signals were returned to the time domain. Both settings used the same agent population, task process, poisoning intensity, and number of update cycles. This design allowed direct comparison between normal synchronization under poisoning and synchronization with spectral containment. The comparison was necessary because poisoning in collaborative systems often spreads through repeated signal exchange, and a method that reduces abnormal spectral components may lower later contamination without fully removing useful coordination content.

2.3. Measurement Method and Quality Control

During each update cycle, memory transition signals from all agents were collected and arranged as temporal sequences for spectral analysis. Discrete Fourier transformation was then applied to each signal to estimate its frequency components and to measure spectral energy across different frequency bands. High-frequency energy was used as the main indicator of poisoning-related disturbance because adversarial perturbations often introduced sharp local changes in the update sequence. In the filtering setting, abnormal bands were attenuated before synchronization, and the filtered signals were then used for later coordination. The main outcome measures were high-frequency spectral energy, contamination spread across agents, and task performance relative to the clean baseline. To keep the results stable, all runs used the same signal length, sampling interval, poisoning rate, and coordination task. Repeated trials were performed under each condition, and the final results were reported as mean values. Signal logs were checked after each run to ensure that the detected spectral changes came from poisoning rather than from numerical noise or inconsistent simulation settings.

2.4. Data Processing and Model Formulation

The recorded memory update sequences were first normalized and then transformed into the frequency domain for band-wise analysis. Let $x(n)$ denote the memory update signal of length N. Its discrete Fourier transform was defined as

$$X(k)=\sum_{n=0}^{N-1} x(n)e^{-j2\pi kn/N}, \quad k=0,1,\ldots,N-1,$$

where $X(k)$ is the spectral component at frequency index k. Spectral energy in a selected frequency band was calculated as

$$E_b=\sum_{k\in B} |X(k)|^2,$$

where B denotes the target frequency band and E_b is the corresponding band energy. Contamination spread was measured as the proportion of agents whose memory states were affected by poisoned updates after synchronization. Task recovery was evaluated by

e-ISSN: 3079-6377 p-ISSN: 3079-6369

comparing filtered task performance with the nominal clean-task performance. All outputs were grouped by condition and summarized across repeated trials.

2.5. Evaluation Criteria and Statistical Analysis

The proposed method was evaluated from three aspects: spectral disturbance, contamination control, and task preservation. Spectral disturbance was measured by the increase in high-frequency energy under poisoning relative to the clean baseline. Contamination control was measured by the proportion of agents affected by poisoned memory after synchronization. Task preservation was measured by the percentage of nominal task performance retained after spectral filtering. These indicators were examined together because an effective containment method should not only reduce abnormal propagation, but also preserve useful coordination quality. To reduce random variation, repeated runs were carried out under each setting, and average values were used in the final comparison. Relative changes were reported as percentage increases or decreases between the filtered and unfiltered conditions. This evaluation made it possible to judge whether the proposed frequency-domain framework could identify poisoning-related spectral change and limit its spread without causing major loss of task performance.

3.Results and Discussion

3.1. Spectral changes caused by poisoning

Poisoned memory updates showed a clear change in spectral structure. At a poisoning intensity of 15%, high-frequency spectral energy increased by 63.2% compared with the clean baseline. This result suggests that adversarial perturbations did not act as random noise only. Instead, they introduced sharper local variation into the update signals, which became more visible after Fourier decomposition. This pattern supports the main assumption of the study, namely that poisoning propagation leaves structured traces in the frequency domain [19,20]. Earlier anomaly studies have reported similar findings, showing that frequency information can reveal changes that are less clear in raw sequences, especially when abnormal behavior appears as oscillation or local signal distortion rather than as a large shift in mean level (Fig. 1).

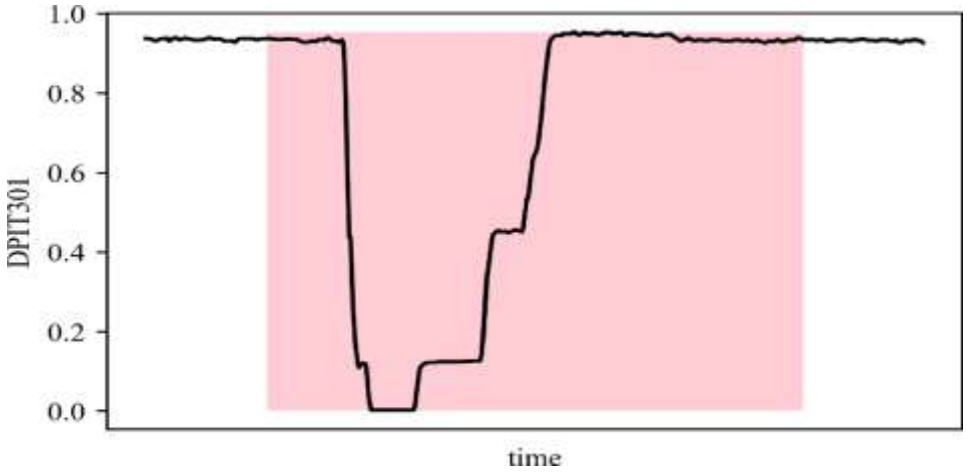

Figure 1:High-frequency spectral energy of memory update signals under clean and poisoned conditions.

e-ISSN: 3079-6377 p-ISSN: 3079-6369

3.2. Effect of spectral filtering on contamination spread

The spectral attenuation filter greatly reduced the spread of contaminated memory. Without filtering, poisoning reached 58.4% of agents. After filtering, the affected proportion dropped to 19.6%. This result shows that the abnormal spectral bands carried a large part of the harmful propagation signal. Once these bands were suppressed before synchronization, later contamination across the agent population became much more limited. The result also shows that containment can be achieved without fully blocking memory exchange. This point matters in collaborative systems, where strong isolation may protect integrity but also damage coordination. In the present case, selective filtering provided a balanced solution: it reduced harmful propagation while keeping enough useful signal for later interaction [21,22]. Recent work on graph-based and frequency-aware anomaly analysis supports this view and suggests that structure-aware filtering can be more effective than direct pointwise suppression in networked systems (Fig. 2).

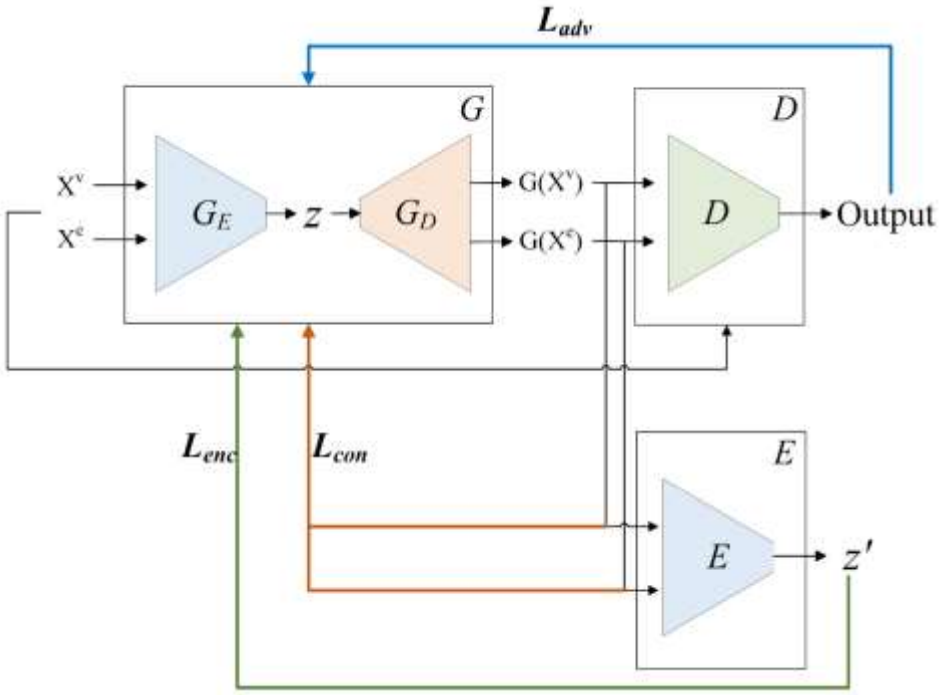

Figure 2:Contamination spread across agents before and after spectral filtering.

3.3. Recovery of task performance after filtering

Task performance also improved after spectral filtering. Under poisoning, filtered coordination restored task performance to 87.3% of the nominal clean level. This result shows that the proposed method did not simply remove suspicious components at the cost of serious information loss. Instead, it preserved most of the useful coordination content while reducing the part of the signal linked to poisoning. This balance is a key strength of the frequency-domain design. In many collaborative systems, a defense method is useful only when it reduces attack impact without harming later task execution [23]. Compared with approaches that focus only on anomaly scoring or attack detection, the present method provides a more direct link between signal analysis and system-level recovery. It therefore extends earlier time-series anomaly work by connecting spectral abnormality with collaborative performance rather than with detection accuracy alone.

e-ISSN: 3079-6377 p-ISSN: 3079-6369

3.4. Comparison with earlier studies and study limits

Compared with earlier studies, the main strength of the present work is that it examines poisoning propagation as a signal-structure problem in a multi-agent memory channel. Many recent studies on multi-agent security focus on reward loss, policy failure, or communication breakdown under attack. Other studies in anomaly detection focus on general multivariate signals, but not on how corrupted memory moves across an interacting agent population. The current results help connect these two research lines by showing that poisoning propagation has a measurable spectral pattern and that this pattern can be used for containment before synchronization. At the same time, the study has several limits. The experiments were carried out in a 120-agent simulation with a fixed poisoning intensity and 10,000 update cycles. Real systems may involve changing topologies, mixed attack forms, and more complex signal overlap between normal coordination and malicious perturbation. Future work should therefore test adaptive spectral filters, broader attack settings, and larger collaborative platforms.

4.Conclusion

This study examined memory poisoning propagation in multi-agent collaborative systems from a frequency-domain perspective. The results showed that poisoned memory updates caused a clear rise in high-frequency spectral energy, with a 63.2% increase over the clean baseline at a poisoning intensity of 15%. After spectral attenuation filtering was applied before synchronization, contamination spread decreased from 58.4% to 19.6% of agents, and task performance recovered to 87.3% of the nominal level. These findings indicate that memory poisoning is not only a task-level disturbance, but also a structured signal pattern that can be identified and controlled in the frequency domain. The main contribution of this work is that it shifts the analysis of poisoning propagation from the time domain to the spectral domain and shows that abnormal frequency components can serve as useful indicators of adversarial diffusion in shared memory channels. This gives the study clear scientific value by linking signal decomposition with collaborative system security and by offering a new way to understand how poisoned updates spread through agent interaction. The proposed framework also has practical value for distributed coordination, cooperative robotics, and other multi-agent platforms in which repeated memory exchange directly affects later system behavior. At the same time, the present study was carried out in a controlled 120-agent simulation with a fixed poisoning intensity and a fixed communication setting. Real systems may involve changing topologies, mixed attack patterns, and more complex overlap between normal coordination signals and malicious perturbations. Future work should therefore examine adaptive spectral filters, broader attack settings, and larger collaborative systems. Overall, the results suggest that frequency-domain modeling is a practical and effective way to detect and contain memory poisoning in collaborative agent environments.

References

[1] Liu, H., Xu, D., Ma, Q., Xu, S., & Qiu, D. (2026). Memory Poisoning Propagation and Repair Mechanism in Multi-Agent Collaborative Environments.

[2] Aslam, M. M., Ahmed, Z., Du, L., Hassan, M. Z., Ali, S., & Nasir, M. (2022). An overview of recent advances of resilient consensus for multiagent systems under attacks. Computational intelligence and neuroscience, 2022(1), 6732343.

e-ISSN: 3079-6377 p-ISSN: 3079-6369

[3] Bai, W., Wu, Q., Wu, K., & Lu, K. (2024). Exploring the Influence of Prompts in LLMs for Security-Related Tasks. In Workshop on Artificial Intelligence System with Confidential Computing (AISCC 2024)(San Diego, CA). USA. https://dx. doi. org/10.14722/aiscc.

[4] Bousetouane, F. (2026). AI Agents Need Memory Control Over More Context. arXiv preprint arXiv:2601.11653.

[5] Du, Y. (2025). Research on Deep Learning Models for Forecasting Cross-Border Trade Demand Driven by Multi-Source Time-Series Data. Journal of Science, Innovation & Social Impact, 1(2), 63-70.

[6] La Gatta, V., Orlando, G. M., Perillo, M., Tammaro, F., & Moscato, V. (2026). From Who They Are to How They Act: Behavioral Traits in Generative Agent-Based Models of Social Media. arXiv preprint arXiv:2601.15114.

[7] Liu, S., Feng, H., & Liu, X. (2025). A Study on the Mechanism of Generative Design Tools' Impact on Visual Language Reconstruction: An Interactive Analysis of Semantic Mapping and User Cognition. Authorea Preprints.

[8] Maleki, S., & Pourmoazemi, N. (2025). Pi-Transformer: A Physics-informed Attention Mechanism for Time Series Anomaly Detection. arXiv preprint arXiv:2509.19985.

[9] Qiu, Y. (2024). Estimation of tail risk measures in finance: Approaches to extreme value mixture modeling. arXiv preprint arXiv:2407.05933.

[10] Keil, A., Bernat, E. M., Cohen, M. X., Ding, M., Fabiani, M., Gratton, G., ... & Weisz, N. (2022). Recommendations and publication guidelines for studies using frequency domain and time-frequency domain analyses of neural time series. Psychophysiology, 59(5), e14052.

[11] Chen, H., Li, J., Ma, X., & Mao, Y. (2025, June). Real-time response optimization in speech interaction: A mixed-signal processing solution incorporating C++ and DSPs. In 2025 7th International Conference on Artificial Intelligence Technologies and Applications (ICAITA) (pp. 110-114). IEEE.

[12] Yahya, M. A., Moya, A. R., & Ventura, S. (2025). Deep learning for multivariate time series anomaly detection: An evaluation of reconstruction-based methods. Artificial Intelligence Review, 58(12), 400.

[13] Li, T., Liu, S., Hong, E., & Xia, J. (2025). Human Resource Optimization in the Hospitality Industry Big Data Forecasting and Cross-Cultural Engagement.

[14] Jaber, M. (2025). Structural and spectral analysis of dynamic graphs for attack detection (Doctoral dissertation, Université de Strasbourg).

[15] Ma, Q., Yue, L., Xu, S., Shi, Y., & Liu, H. (2026). Web Agent Agentic Reinforcement Learning Decision Model Under Multi-Cost and Failure Risk Constraints.

[16] Evangelatos, S., Veroni, E., Konidi, M., Karagiorgou, S., Dede, G., Baklezos, A., ... & Goudos, S. K. (2025). Adaptive policy-oriented cybersecurity: A decentralized framework using message passing algorithms for dynamic threat mitigation. IEEE Access.

[17] Evangelatos, S., Veroni, E., Konidi, M., Karagiorgou, S., Dede, G., Baklezos, A., ... & Goudos, S. K. (2025). Adaptive policy-oriented cybersecurity: A decentralized framework using message passing algorithms for dynamic threat mitigation. IEEE Access.

[18] Qiu, D., Xu, D., & Yue, L. (2025, December). Reinforcement Learning-Augmented LLM Agents for Collaborative Decision Making and Performance Optimization. In 2025 7th International Conference on Frontier Technologies of Information and Computer (ICFTIC) (pp. 1337-1342). IEEE.

e-ISSN: 3079-6377 p-ISSN: 3079-6369

[19] Rawat, R., Rawat, H., & Rawat, A. (2025). Energy-aware detection of threat information propagation speed in social network of things using XGBoost and sequential pattern mining. Discover Computing, 28(1), 223.

[20] Gu, X., Yang, J., Tian, X., & Liu, M. (2025). Research on the Construction of a Human-Machine Collaborative Anti-Money Laundering System and Its Efficiency and Accuracy Enhancement in Suspicious Transaction Identification.

[21] Rouholamini, S. R., Mirabi, M., Farazkish, R., & Sahafi, A. (2024). Proactive self-healing techniques for cloud computing: A systematic review. Concurrency and Computation: Practice and Experience, 36(24), e8246.

[22] Yang, Y., Leuze, C., Hargreaves, B., Daniel, B., & Baik, F. (2025). EasyREG: Easy Depth-Based Markerless Registration and Tracking using Augmented Reality Device for Surgical Guidance. arXiv preprint arXiv:2504.09498.

[23] Rai, M., & Kesarwani, A. (2025). Image watermarking in the digital era: A review of transform, ML-based, and evolutionary methods. Circuits, Systems, and Signal Processing, 1-52.