Novelty Assessment Report

Paper: Remotely Detectable Robot Policy Watermarking
PDF URL: https://openreview.net/pdf?id=8s5jBVybhQ
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-30

Abstract

The success of machine learning for real-world robotic systems has created a new form of intellectual property: the trained policy. This raises a critical need for novel methods that verify ownership and detect unauthorized, possibly unsafe misuse. While watermarking is established in other domains, physical policies present a unique challenge: remote detection. Existing methods assume access to the robot's internal state, but auditors are often limited to external observations (e.g., video footage). This "Physical Observation Gap" means the watermark must be detected from signals that are noisy, asynchronous, and filtered by unknown system dynamics. We formalize this challenge using the concept of a glimpse sequence, and introduce Colored Noise Coherency (CoNoCo), the first watermarking strategy designed for remote detection. CoNoCo embeds a spectral signal into the robot's motions by leveraging the policy's inherent stochasticity. To show it does not degrade performance, we prove CoNoCo preserves the marginal action distribution. Our experiments demonstrate strong, robust detection across various remote modalities—including motion capture and side-way/top-down video footage—in both simulated and real-world robot experiments. This work provides a necessary step toward protecting intellectual property in robotics, offering the first method for validating the provenance of physical policies non invasively, using purely remote observations.

Core Task Landscape

This paper addresses: **Remotely Detectable Robot Policy Watermarking Using Frequency Domain Analysis**

A total of **1 papers** were analyzed and organized into a taxonomy with **2 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Frequency-Based Watermarking for Robot Policies**

Complete Taxonomy Tree

- Remotely Detectable Robot Policy Watermarking Using Frequency Domain Analysis Survey Taxonomy
- Frequency-Based Watermarking for Robot Policies
  - Colored Noise Coherency Watermarking ★ (1 papers)
  - [0] Remotely Detectable Robot Policy Watermarking (Anon et al., 2026) View paper
  - Frequency-Based Replay Attack Detection (1 papers)
  - [1] KHALIL GUIBENE (N MESSAI, n.d.) View paper

Narrative

Core task: remotely detectable robot policy watermarking using frequency domain analysis. This emerging area addresses the challenge of verifying ownership or authenticity of deployed robot policies by embedding detectable signatures in their behavior. The taxonomy centers on frequency-based watermarking techniques, which exploit the spectral properties of control signals or trajectories to insert imperceptible markers. Within this single top-level branch, the field explores methods that modulate robot actions in the frequency domain—such as injecting colored noise patterns—so that an observer can later extract and verify the watermark without direct access to the policy's internal parameters. Representative works like Robot Policy Watermarking[0] demonstrate how coherency analysis of frequency components can reveal embedded signatures while maintaining task performance.

A key theme across this line of work is the trade-off between watermark robustness and the subtlety of behavioral perturbations: stronger frequency signatures improve detectability but risk degrading control quality or becoming noticeable to adversaries. Robot Policy Watermarking[0] sits within the colored noise coherency watermarking cluster, emphasizing spectral coherence measures to achieve remote detection without requiring model access. This approach contrasts with potential alternatives that might embed watermarks in time-domain statistics or rely on cryptographic hashes of policy weights. The focus on frequency domain analysis reflects a broader interest in leveraging signal processing tools to balance imperceptibility, verifiability, and resilience against policy modifications or adversarial removal attempts.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

Both subtopics apply frequency domain analysis to robot policy watermarking, but serve fundamentally different purposes. The original leaf focuses on embedding detectable watermarks into policy stochasticity for ownership verification, while the sibling focuses on using watermarks to detect replay attacks as a security mechanism.

Similarities: - Both leverage frequency domain analysis as the core technical approach - Both operate on robotic systems and policies - Both enable remote detection without requiring internal state access - Both use spectral characteristics as the basis for their detection mechanisms

Differences: - Purpose: Original leaf is for policy ownership/provenance verification; sibling is for security threat detection (replay attacks) - Application context: Original leaf applies to general policy stochasticity watermarking; sibling specifically targets robotic arms with multiple degrees of freedom - Detection target: Original leaf detects presence of embedded watermark signatures; sibling detects

anomalous replay patterns - Threat model: Original leaf assumes benign deployment with ownership disputes; sibling assumes active adversarial replay attacks - Noise type: Original leaf explicitly uses colored noise coherency for watermark embedding; sibling uses frequency analysis for anomaly detection

Suggested Search Directions: - Hybrid approaches combining ownership watermarking with security threat detection in the frequency domain - Adversarial robustness of frequency-based watermarks against replay attacks - Unified frameworks for multi-purpose frequency domain analysis in robot policies (authentication, security, provenance)

Sibling Subtopics

- **Frequency-Based Replay Attack Detection** (leaves: 1, papers: 1)
- Scope: Watermark-based detection methods using frequency analysis to identify replay attacks on robotic arms with multiple degrees of freedom.
- Exclude: Non-replay attack scenarios and non-frequency detection methods belong to other security or watermarking categories.

Contributions Analysis

Overall novelty summary. The paper introduces a watermarking framework for robot policies that enables remote detection through external observations, addressing what the authors term the 'Physical Observation Gap.' Within the taxonomy, this work occupies the 'Colored Noise Coherency Watermarking' leaf under 'Frequency-Based Watermarking for Robot Policies.' Notably, this leaf contains only the original paper itself, with no sibling papers identified. The taxonomy as a whole comprises just two papers across two leaves, suggesting this is an emerging and sparsely populated research direction rather than a crowded subfield.

The taxonomy structure reveals that the broader category of 'Frequency-Based Watermarking for Robot Policies' contains one neighboring leaf focused on 'Frequency-Based Replay Attack Detection,' which addresses security concerns in robotic arms using frequency analysis. This neighboring work targets a different problem (replay attacks) and application context (robotic arms with multiple degrees of freedom), while the original paper focuses on ownership verification and misuse detection across general robotic systems. The taxonomy's scope notes clarify that non-frequency watermarking approaches and non-robotic watermarking fall outside this branch, positioning the work within a specific intersection of signal processing and robot policy protection.

Among the three contributions analyzed, the literature search examined only one candidate paper total, finding no clear refutations for any contribution. Specifically, the theoretical guarantee of marginal action distribution preservation was examined against one candidate, which was classified as non-refutable or unclear. The formalization of glimpse sequences and the CoNoCo strategy itself were examined against zero candidates. Given this extremely limited search scope—one candidate paper across all contributions—the analysis provides minimal evidence about prior work overlap. The absence of refutable candidates may reflect either genuine novelty or insufficient literature coverage.

Based on the single-paper taxonomy and minimal literature search (one candidate examined), the work appears to occupy a nascent research area with limited documented prior art. However, the analysis explicitly acknowledges its scope limitations: the search was not exhaustive, relying on top-K semantic matching. The sparse taxonomy and zero-sibling-paper finding suggest either that this specific formulation is genuinely novel or that related work exists under different terminology or in adjacent communities not captured by the search methodology.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Formalization of remotely detectable policy watermarking using glimpse sequences

Description: The authors introduce a formal framework for robot policy watermarking that must be detected from remote observations only. They define glimpse sequences to model the Physical Observation Gap and identify three core challenges: synchronization uncertainty, system dynamics filtering, and interference plus noise.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 2: Colored Noise Coherency (CoNoCo) watermarking strategy

Description: The authors propose CoNoCo, a watermarking method that embeds spectral signatures by replacing white Gaussian noise with colored Gaussian noise in the policy's exploration, and detects these signatures using spectral coherency. This approach is designed specifically to enable remote detection despite unknown system dynamics and asynchronous sensing.

This contribution was assessed against **0 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

Contribution 3: Theoretical guarantee of marginal action distribution preservation

Description: The authors provide a theoretical proof (Theorem 4.1) demonstrating that their watermarking approach preserves the statistical distribution of actions at any single time step, ensuring the watermarked policy behaves identically to the original policy in terms of marginal action probabilities.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Watermarks for Deep Reinforcement Learning

URL: View paper

Brief Assessment

Deep RL Watermarks[2] focuses on watermarking stochastic DRL policies but does not provide theoretical guarantees about preserving marginal action distributions. The candidate's context only mentions that 'stochastic drl policies choose actions in different ways' without addressing distribution preservation theory.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Remotely Detectable Robot Policy Watermarking View paper
- [1] KHALIL GUIBENE View paper
- [2] Watermarks for Deep Reinforcement Learning View paper