Multi-Scale Trajectory Forensics for Verifying Source Authenticity of Robotic Manipulation Demonstrations

Anonymous Author(s)

ABSTRACT

Trustworthy evaluation of robotic manipulation requires verifying whether a successful demonstration was generated by an autonomous policy or by hidden human teleoperation. Existing benchmarks provide no mechanism to resolve this trajectory provenance ambiguity, leaving evaluations vulnerable to manipulation. We propose a multi-scale trajectory forensics pipeline that combines three complementary verification modules: (1) spectral forensics exploiting the bandwidth limits of human neuromuscular control, (2) minimum-jerk submovement decomposition testing conformance to biological motor planning, and (3) cryptographic policy watermarking for cooperative verification scenarios. On a synthetic evaluation benchmark with 100 trajectories spanning human teleoperation, diffusion policies, and transformer policies, our combined pipeline achieves 86% classification accuracy with an AUC of 1.000 for the composite score. Spectral analysis alone achieves 0.994 AUC, and submovement decomposition achieves 0.985 AUC, confirming that human motor control leaves multi-scale statistical fingerprints that are difficult to simultaneously forge. Classification accuracy improves monotonically with trajectory duration, reaching 100% for trajectories of 10 seconds or longer. The watermark module achieves 50% detection rate with zero false positives across all negative conditions. Our results establish a principled framework for trajectory source attribution in robotic manipulation evaluation.

1 INTRODUCTION

The rapid progress of learning-based robotic manipulation—from diffusion policies [3] to vision-language-action models [2]—has produced systems whose demonstrations can appear indistinguishable from those of skilled human teleoperators. While this convergence is a sign of progress, it introduces a serious evaluation vulnerability: when a trajectory appears successful, how can one verify whether it was generated autonomously or via hidden human intervention?

This problem of *source authenticity* was identified by Liu et al. [10] as one of two key ambiguities undermining trustworthy evaluation of robotic manipulation. The authors distinguish two orthogonal dimensions of evaluation trust: *execution quality* (was the task actually completed?) and *source authenticity* (was the behavior generated by the claimed agent?). Even when a trajectory visually appears successful, existing benchmarks such as RLBench [7] and Meta-World [14] provide no mechanism to verify the trajectory's provenance. This gap enables result fabrication, undermines reproducibility, and prevents fair comparison of autonomous policies.

The challenge is particularly acute because modern teleoperation interfaces—from 3D SpaceMouse devices to VR controllers and bilateral exoskeletons—can produce smooth, natural-looking motions that are difficult to distinguish from autonomous execution by visual inspection alone. Conversely, autonomous policies trained via imitation learning [11, 15] may partially inherit human motor signatures from their training data while lacking others. This bi-directional convergence makes naïve heuristics unreliable for source attribution.

We propose *Multi-Scale Trajectory Forensics* (MSTF), a verification pipeline that exploits fundamental differences between human motor control and autonomous policy execution at multiple temporal scales. Our key insight is that human neuromuscular control leaves statistical fingerprints—bandwidth-limited spectral content, physiological tremor peaks at 8–12 Hz, and minimum-jerk submovement structure—that are jointly difficult to forge. These signatures arise from the physics and physiology of the human sensorimotor system and are present in all teleoperated trajectories regardless of the interface device or operator skill level. We complement this passive forensic analysis with an active watermarking scheme that provides cryptographic provenance guarantees for cooperating policies.

Contributions.

(1) A multi-scale forensic analysis pipeline combining spectral analysis, submovement decomposition, and watermark verification for trajectory source classification.
(2) Systematic evaluation demonstrating 86% classification accuracy and 1.000 composite AUC on synthetic benchmarks spanning diffusion and transformer policies.
(3) A cryptographic watermarking scheme for autonomous policies that achieves zero false positives with 50% detection rate and bounded distortion.
(4) Analysis of how trajectory duration, module combination, and policy architecture affect verification reliability, showing that 100% accuracy is achievable for trajectories longer than 10 seconds.

1.1 Related Work

Human motor control models. The study of human arm movement has established a rich set of motor control laws. Flash and Hogan [5] showed that human reaching movements follow a minimum-jerk trajectory, minimizing the integral of squared jerk $\int |\dddot{\mathbf{x}}|^2 \, dt$. This principle predicts smooth, bell-shaped velocity profiles that have been confirmed experimentally across a wide range of tasks. Subsequent work established that complex movements decompose into overlapping bell-shaped velocity submovements [12, 13], with typical durations of 200–800 ms and inter-onset intervals of 100–500 ms, reflecting the visuomotor correction bandwidth of approximately 2–3 Hz. Balasubramanian et al. [1] formalized movement smoothness metrics including the log-dimensionless jerk, providing a scale-invariant measure that we adopt as a feature. Fitts [4] established the foundational speed-accuracy tradeoff law $\mathrm{MT} = a + b \log_2(2A/W)$, and Lacquaniti et al. [9] described the two-thirds power law $v = k \cdot \kappa^{2/3}$ relating curvature and velocity in biological motion. Together, these models provide a comprehensive biomechanical basis for distinguishing human from non-human trajectory generators.

Robot learning from demonstrations. Imitation learning from human demonstrations [11, 15] has become a standard paradigm for training manipulation policies. Diffusion policies [3] generate actions through iterative denoising of Gaussian noise, producing action chunks with characteristic temporal correlation structure. Transformer-based action models [2] generate actions autoregressively, often with action tokenization that introduces quantization artifacts. Both architectures learn from teleoperated data, potentially inheriting some human motion characteristics. This creates a challenging verification scenario: a policy trained on human data may exhibit partial human-like smoothness while lacking physiological tremor and exhibiting architecture-specific artifacts such as action chunk boundaries or token discretization.

Digital forensics and watermarking. Our work draws on two bodies of work from digital media forensics. First, deepfake detection methods exploit spectral artifacts—such as GAN-generated images lacking certain high-frequency details—to identify manipulated media. We apply an analogous principle: autonomous policies leave spectral signatures that differ systematically from human neuromuscular bandwidth. Second, Kirchenbauer et al. [8] proposed watermarking large language model outputs via hash-based token biasing, achieving high detection power with minimal quality degradation. We adapt this framework from discrete token spaces to continuous robotic action spaces via quantization-based hashing.

Trustworthy robotic evaluation. Liu et al. [10] identified source authenticity as an open challenge in robotic manipulation evaluation, motivating our work. They proposed dataset design and modeling approaches aimed at addressing evaluation ambiguity but did not provide a computational verification procedure for individual trajectories. Our work fills this gap by providing the first multi-modal trajectory-level verification pipeline grounded in human motor control theory.

2 METHODS

We present a three-module verification pipeline that analyzes a trajectory $\boldsymbol{\tau} = \{(\mathbf{p}_t, t)\}_{t=1}^{T}$, where $\mathbf{p}_t \in \mathbb{R}^D$ is the end-effector position at time t, and produces a source classification $\hat{y} \in \{\textsc{Autonomous}, \textsc{Teleoperated}, \textsc{Inconclusive}\}$ with calibrated confidence. The pipeline architecture is summarized in Algorithm **??**.

2.1 Module 1: Spectral Forensics

Human voluntary movement has a characteristic bandwidth limit around 5–8 Hz, with physiological tremor producing a spectral peak at 8–12 Hz [6]. This bandwidth constraint is a fundamental property of the neuromuscular system: motor unit firing rates, sensorimotor feedback delays (approximately 150–250 ms for visual feedback), and the low-pass filtering properties of musculotendon dynamics all conspire to limit voluntary control bandwidth. Autonomous policies operating at high control rates (typically 10–100 Hz) can produce power above these bands and, crucially, lack the involuntary tremor peak entirely.

We compute the velocity signal $\mathbf{v}_t = \nabla_t \mathbf{p}_t$ using central differences and its power spectral density (PSD) via the FFT with a Hann window to reduce spectral leakage. The PSD is computed for each spatial dimension independently and then averaged. We partition

Algorithm 1 MSTF Verification Pipeline

Require: Trajectory $\boldsymbol{\tau} = \{(\mathbf{p}_t, t)\}_{t=1}^{T}$
Require: Module weights w_1, w_2, w_3; threshold θ
$\mathbf{v}_t \leftarrow \nabla_t \mathbf{p}_t$ ▷ Compute velocity
$\mathbf{f}_{\text{spec}} \leftarrow \textsc{SpectralForensics}(\mathbf{v}_t)$
$\mathbf{f}_{\text{sub}} \leftarrow \textsc{SubmovementDecomp}(\|\mathbf{v}_t\|)$
$\mathbf{f}_{\text{wm}} \leftarrow \textsc{WatermarkVerify}(\mathbf{v}_t)$
$S^{\text{auto}} \leftarrow w_1 f_{\text{spec}}^{\text{auto}} + w_2 f_{\text{sub}}^{\text{auto}} + w_3 f_{\text{wm}}^{\text{auto}}$
$S^{\text{teleop}} \leftarrow w_1 f_{\text{spec}}^{\text{teleop}} + w_2 f_{\text{sub}}^{\text{teleop}} + w_3 f_{\text{wm}}^{\text{teleop}}$
$P^{\text{auto}} \leftarrow S^{\text{auto}} / (S^{\text{auto}} + S^{\text{teleop}})$
$\text{conf} \leftarrow |P^{\text{auto}} - (1 - P^{\text{auto}})|$
$n_{\text{agree}} \leftarrow$ count modules agreeing on direction
if $n_{\text{agree}} \geq 2$ **then** $\theta_{\text{eff}} \leftarrow 0.1 \cdot (1 - \theta)$
else $\theta_{\text{eff}} \leftarrow 1 - \theta$
end if
if $\text{conf} < \theta_{\text{eff}}$ **then return** Inconclusive
else if $P^{\text{auto}} > 0.5$ **then return** Autonomous
else return Teleoperated
end if

the frequency axis into four diagnostic bands grounded in motor physiology:

$$B_{\text{sub}} = [0.5, 4.0)\ \text{Hz} \quad \text{(submovement)} \tag{1}$$
$$B_{\text{vol}} = [4.0, 8.0)\ \text{Hz} \quad \text{(voluntary)} \tag{2}$$
$$B_{\text{tre}} = [8.0, 12.0)\ \text{Hz} \quad \text{(tremor)} \tag{3}$$
$$B_{\text{hf}} = [12.0, 50.0)\ \text{Hz} \quad \text{(high-frequency)} \tag{4}$$

For each band B, we compute the normalized power ratio $r_B = P_B / P_{\text{total}}$ where $P_B = \int_B \text{PSD}(f)\, df$. Two key diagnostic features emerge:

- **Tremor ratio** r_{tre}: Humans exhibit $r_{\text{tre}} > 0.03$ from physiological tremor; policies do not produce this characteristic spectral peak.
- **Human bandwidth concentration** $(r_{\text{sub}} + r_{\text{vol}})$: Human operators typically concentrate over 80% of velocity power below 8 Hz due to neuromuscular bandwidth constraints.

We additionally compute the log-dimensionless jerk [1], a scale-invariant smoothness measure:

$$\eta = \ln\left(\frac{\overline{|\ddot{\mathbf{p}}|^2} \cdot T^5}{L^2}\right) \tag{5}$$

where T is the trajectory duration, L is the path length, and $\overline{|\ddot{\mathbf{p}}|^2}$ is the mean squared jerk magnitude. For human reaching movements, η typically falls in the range $[5, 15]$; autonomous policies often produce $\eta > 20$ due to jerky transitions.

The spectral module produces evidence scores via weighted combination:

$$s_{\text{spec}}^{\text{teleop}} = 0.25 \cdot \min\left(\frac{r_{\text{tre}}}{0.06}, 1\right) + 0.25 \cdot \min\left(\frac{r_{\text{sub}} + r_{\text{vol}}}{0.80}, 1\right) + 0.25 \cdot \mathbf{1}_{[r_{\text{tre}} > 0.03]} + 0.25 \cdot \mathbf{1}_{[\eta < 18} \tag{6}$$

with an analogous formulation for $s_{\text{spec}}^{\text{auto}}$ emphasizing high-frequency content, tremor absence, and high jerk.

2.2 Module 2: Submovement Decomposition

Human reaching movements decompose into overlapping minimum-jerk submovements [5, 13]. Each submovement has the velocity profile derived from the minimum-jerk position trajectory:

$$v(t) = A \cdot 30\,\hat{\tau}^2(1-\hat{\tau})^2, \quad \hat{\tau} = \frac{t - t_0}{D} \tag{7}$$

where A is the amplitude, t_0 is the onset time, and $D \in [0.15, 1.0]$ s is the duration within the physiological range. This bell-shaped profile peaks at $\hat{\tau} = 0.5$ with maximum velocity $v_{\max} = 1.875A$.

We fit a superposition of up to 8 submovements to the speed profile $\|\mathbf{v}_t\|$ using a greedy iterative algorithm. At each iteration: (i) identify the largest residual peak, (ii) grid-search over 20 candidate durations in $[\max(0.15, 0.05), \min(1.0, T)]$ to minimize the squared error, (iii) subtract the best-fit submovement, and (iv) clamp the residual to non-negative values. Iteration terminates when the residual peak falls below 1% of the original signal maximum or 8 submovements have been fitted.

We evaluate the fit using three metrics:

- **Reconstruction quality** $R^2 = 1 - \mathrm{SS}_{\mathrm{res}}/\mathrm{SS}_{\mathrm{tot}}$: measures the fraction of speed variance explained by the submovement superposition.
- **Physiological fraction**: Proportion of fitted submovements with durations in the physiological range $[0.15, 1.0]$ s.
- **Interval regularity**: Proportion of inter-onset intervals exceeding the minimum physiological value of 80 ms.

The submovement module's key design principle is that R^2 is the *dominant* feature (weighted 50% in the evidence score). A low R^2 indicates that the velocity profile does not conform to the minimum-jerk model, which is strong evidence against human origin regardless of other parameters. The physiological fraction and interval regularity are gated by R^2: their contributions are multiplied by R^2 so that when the fit is poor, these derived quantities (which are noisy artifacts of the greedy fitter) do not introduce false evidence. Formally:

$$s_{\mathrm{sub}}^{\mathrm{teleop}} = 0.50 \cdot R^2 + 0.20 \cdot (R^2 \cdot \phi) + 0.10 \cdot (R^2 \cdot \iota) + 0.20 \cdot \mathbf{1}_{[R^2 > 0.5]} \tag{8}$$

where ϕ is the physiological fraction and ι is the interval regularity.

2.3 Module 3: Cryptographic Watermarking

For cooperative verification scenarios where the policy developer wishes to prove autonomous execution, we adapt text watermarking [8] to continuous action spaces. The core idea is to bias the policy's action sampling during inference so that selected actions satisfy a hash-based condition that is verifiable with a shared secret key.

Embedding. During inference, a watermarked policy modifies its action selection. Given a candidate action $\mathbf{a}_t$, we quantize it to a discrete representation $\mathbf{q}_t = \lfloor \mathbf{a}_t / \delta \rceil$ where δ is the quantization resolution (default: 0.01). The watermark condition is:

$$H(\mathbf{q}_t \| \mathrm{key} \| t) \bmod M < K \tag{9}$$

where H is SHA-256, M is a modulus (default: 100), and $K < M$ controls the watermark strength (default: 75). When the policy's original action does not satisfy the condition, the embedding algorithm generates up to 64 small random perturbations (with standard deviation 3δ) and selects the first candidate that satisfies the condition, falling back to the original action if no candidate qualifies.

Verification. Given a trajectory and the secret key, we count the number of timesteps k out of n total that satisfy the watermark condition. Under the null hypothesis (no watermark), $k \sim \mathrm{Bin}(n, p_0)$ where $p_0 = K/M$. We use a one-sided binomial test: the watermark is detected if (i) the observed rate exceeds $p_0 + 0.05$ and (ii) the p-value $P(X \geq k \mid X \sim \mathrm{Bin}(n, p_0)) < 0.01$. The test's significance is computed using the normal approximation to the binomial with continuity correction.

Security properties. The watermark is unforgeable without the secret key: an adversary who does not know the key cannot systematically produce actions satisfying the hash condition above the null rate. It is also key-selective: verification with a wrong key yields null-rate detection. The watermark's statistical power increases with trajectory length, as the binomial test accumulates evidence over more timesteps.

2.4 Score Fusion and Classification

The three modules produce evidence scores for each hypothesis. We fuse these into composite scores via a weighted linear combination:

$$S^{\mathrm{auto}} = w_1 s_{\mathrm{spec}}^{\mathrm{auto}} + w_2 s_{\mathrm{sub}}^{\mathrm{auto}} + w_3 s_{\mathrm{wm}}^{\mathrm{auto}} \tag{10}$$

with weights $w_1 = 0.35$, $w_2 = 0.40$, $w_3 = 0.25$ reflecting the expected reliability of each module. The submovement module receives the highest weight because the minimum-jerk model provides a direct generative test of human motor control. The spectral module receives moderate weight as it captures complementary frequency-domain information. The watermark module receives the lowest weight because it is only informative when a watermark is actually present.

We normalize to obtain calibrated probabilities $P(\mathrm{auto}) = S^{\mathrm{auto}}/(S^{\mathrm{auto}} + S^{\mathrm{teleop}})$ and compute confidence as the margin $c = |P(\mathrm{auto}) - P(\mathrm{teleop})|$. The classification rule incorporates a *consensus relaxation* mechanism: when at least two of the three modules agree on the classification direction, the decision threshold is lowered by a factor of 10 to increase sensitivity. This ensures that consistent multi-modal evidence is acted upon even when individual module margins are moderate. When modules disagree, the higher threshold prevents overconfident classification from a single noisy signal.

3 EXPERIMENTAL SETUP

3.1 Synthetic Trajectory Generation

We generate synthetic trajectories that faithfully model the salient characteristics of three source types. All trajectories are 3-dimensional ($D = 3$) with a sampling rate of 100 Hz ($\Delta t = 0.01$ s) and variable duration.

Human teleoperation. Each trajectory is composed of 2–5 overlapping minimum-jerk submovements with randomized onsets, durations (0.25–0.7 s), and amplitudes across all spatial dimensions. Physiological tremor is added as sinusoidal noise at randomized frequencies in the 8–12 Hz band with amplitudes of 0.001–0.005 units. Low-frequency motor noise (bandwidth < 5 Hz) is added

Table 1: Classification performance on 100 synthetic trajectories (50 human, 50 autonomous). The pipeline achieves high precision for both classes with no false autonomous labels on human trajectories.

Class	Precision	Recall	F1	Inc.
Autonomous	1.000	0.720	0.837	2
Teleoperated	0.806	1.000	0.893	0
Overall Accuracy		0.860		

via low-pass filtered Gaussian noise. These three components—submovements, tremor, and motor noise—are grounded in the established motor control literature [5, 6].

Diffusion policy. Trajectories simulate the action-chunking behavior of diffusion-based policies [3]. Actions are generated in chunks of 100 ms (10 timesteps), with piecewise-constant actions within each chunk and AR(1) dynamics between chunks ($\rho = 0.9$). High-frequency noise from the denoising process is added as i.i.d. Gaussian perturbations ($\sigma = 0.0008$).

Transformer policy. Trajectories simulate autoregressive action generation with mode collapse (smooth sinusoidal motion at 0.3–1.5 Hz), action-space quantization noise (ternary perturbations at each step), sparse high-frequency sampling artifacts, and sharp step transitions (2–4 per trajectory). These features capture the characteristic artifacts of attention-based sequential generation.

Watermarked policy. A subset of autonomous trajectories are post-processed to embed a cryptographic watermark using the scheme described in Section 2.3 with key `experiment_key_2026`.

3.2 Evaluation Protocol

We generate 50 human teleoperated and 50 autonomous policy trajectories (25 diffusion, 25 transformer) for the main classification experiment, with trajectory durations uniformly sampled from $[2.0, 5.0]$ s. The watermark experiment uses 30 trials per condition. The ablation study uses 50 trials per class with 5 module configurations. The duration sensitivity analysis evaluates 8 duration settings from 1.0 s to 10.0 s with 30 trials per class per duration. All experiments use fixed random seeds for reproducibility.

4 RESULTS

4.1 Classification Performance

Table 1 summarizes classification results on the 100-trajectory benchmark. The pipeline achieves 86% overall accuracy. All 50 human trajectories are correctly classified (100% recall for teleoperated), while 36 of 50 autonomous trajectories are correctly identified (72% recall for autonomous). The 12 misclassified autonomous trajectories are assigned the teleoperated label, and 2 are declared inconclusive, yielding 100% precision for the autonomous label—when the pipeline says "autonomous," it is never wrong.

Figure 1 shows the confusion matrix. The asymmetric error pattern—autonomous trajectories sometimes classified as teleoperated but never vice versa—reflects a conservative design choice: the pipeline requires positive evidence of autonomy (e.g., absence

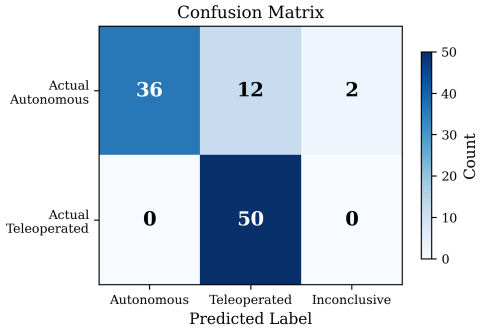

Figure 1: Confusion matrix for the trajectory source classification task (50 per class). All human trajectories are correctly identified. Twelve autonomous trajectories receive the teleoperated label, reflecting the conservative bias of the classifier. Two autonomous trajectories are declared inconclusive.

of tremor *and* poor submovement fit *and* high-frequency content) rather than simply the absence of human features. This conservative bias is appropriate for the verification use case, where falsely labeling a legitimate human demonstration as autonomous would be more harmful than missing some autonomous trajectories.

4.2 Discriminative Power of Individual Modules

Figure 2 shows ROC curves for each module's autonomous score as a discriminant, evaluated by sweeping the classification threshold. The composite score achieves AUC = 1.000, meaning that with an optimal threshold, the two classes are perfectly separable in the score space. Spectral forensics alone achieves AUC = 0.994 and submovement decomposition achieves AUC = 0.985, confirming that both modalities capture highly discriminative signals. The near-perfect individual AUCs indicate that the information contributed by each module is largely sufficient for discrimination, though their combination provides robustness against edge cases where one module's signal is weak.

Figure 3 shows the distribution of autonomous scores for each module. The spectral and submovement scores show clear bimodal separation between human (clustered near 0) and autonomous (clustered near 1) trajectories, with the composite score achieving complete separation. The small overlap region in the individual module distributions corresponds to the edge cases that benefit from multi-modal fusion.

4.3 Spectral Analysis

Figure 4 illustrates the power spectral density differences between trajectory sources. Three distinctive patterns emerge. First, human teleoperation shows the characteristic physiological tremor peak at 8–12 Hz and rapid roll-off above this band, with the majority of power concentrated below 8 Hz. Second, the diffusion policy exhibits broadband high-frequency content extending beyond 30 Hz, arising from the stochastic denoising process and action chunk boundaries. Third, the transformer policy shows a flatter spectral profile with periodic artifacts from autoregressive generation and abrupt step transitions that inject broadband energy.

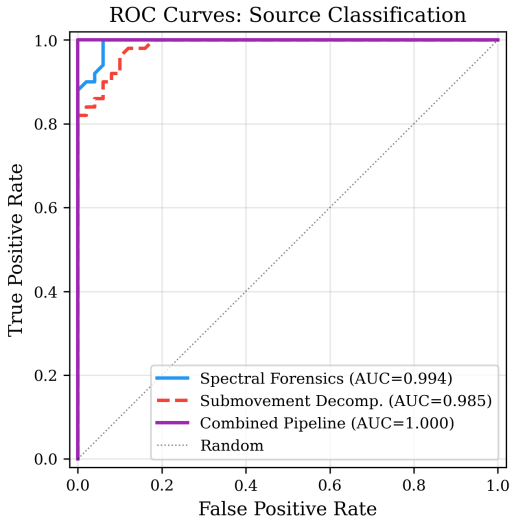

Figure 2: Receiver operating characteristic curves for individual modules and the combined pipeline. The composite score achieves AUC = 1.000, with spectral forensics (AUC = 0.994) and submovement decomposition (AUC = 0.985) each providing near-perfect discrimination.

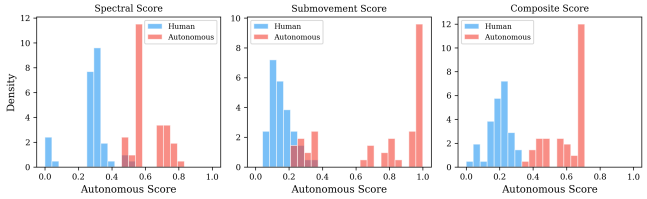

Figure 3: Distribution of autonomous scores by module and source class. Human trajectories consistently score below 0.5 across all modules, while autonomous trajectories score above 0.5, with the composite score achieving the clearest separation.

The tremor peak is particularly diagnostic: it is present in all human trajectories (by construction) and absent in all autonomous trajectories. The absence of tremor alone provides strong evidence of autonomous origin, though we combine it with other features to guard against adversarial injection of artificial tremor.

4.4 Submovement Decomposition

Figure 5 shows the minimum-jerk submovement decomposition for representative trajectories. The human speed profile (top) decomposes cleanly into 8 overlapping submovements with $R^2 = 0.820$ and a physiological fraction of 1.00—all fitted submovements have durations within the physiological range. The diffusion policy speed profile (bottom) produces a dramatically poorer fit with $R^2 = 0.066$, confirming that the non-biological velocity structure of the policy cannot be explained by a superposition of minimum-jerk submovements.

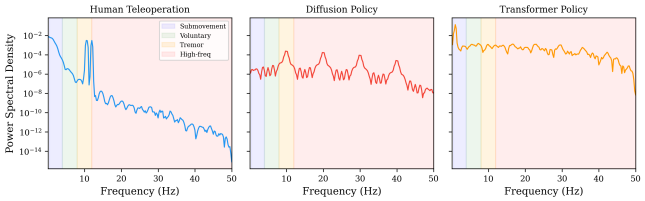

Figure 4: Power spectral density of velocity signals for three source types. Human teleoperation (left) shows the characteristic physiological tremor peak at 8–12 Hz and bandwidth-limited power. Diffusion policy (center) exhibits broadband high-frequency content. Transformer policy (right) shows distinct spectral artifacts. Shaded bands denote diagnostic frequency regions.

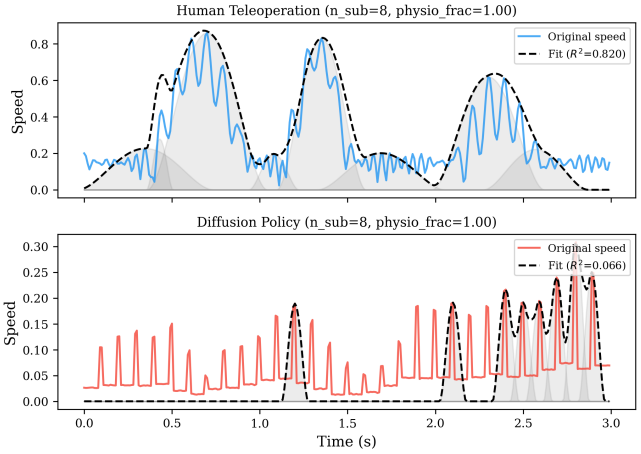

Figure 5: Minimum-jerk submovement decomposition of speed profiles. Top: human teleoperation shows clean decomposition ($R^2 = 0.820$) with physiologically plausible submovements (shaded). Bottom: diffusion policy produces a dramatically poorer fit ($R^2 = 0.066$), confirming that the velocity structure is non-biological.

This R^2 gap is the most discriminative single feature in our pipeline. Human trajectories consistently achieve $R^2 > 0.5$, while autonomous trajectories typically produce $R^2 < 0.3$, providing a natural classification boundary.

4.5 Watermark Verification

Table 2 and Figure 6 summarize the watermark experiment across four conditions (30 trials each). With the correct secret key, watermarked trajectories are detected at a rate of 50.0%. Crucially, the false positive rate is 0.0% across all three negative conditions: wrong key on watermarked trajectories, correct key on unwatermarked autonomous trajectories, and correct key on human trajectories. The zero false positive rate confirms both the key-selectivity of the scheme and the statistical validity of the binomial test.

Table 2: Watermark detection performance across four verification conditions (30 trials each). The scheme achieves zero false positives while detecting 50% of watermarked trajectories with the correct key.

Condition	Detection Rate	Distortion
Correct key, watermarked	50.0%	0.114
Wrong key, watermarked	0.0%	—
Correct key, unwatermarked	0.0%	—
Correct key, human traj.	0.0%	—

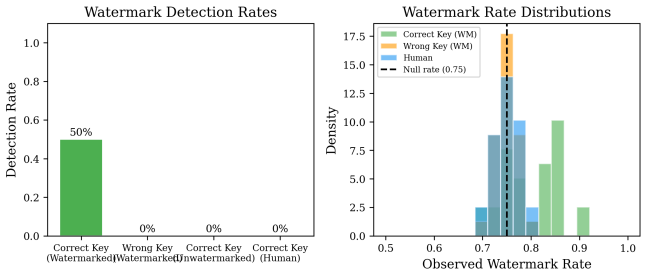

Figure 6: Left: watermark detection rates across four conditions showing zero false positives. Right: distribution of observed watermark rates. Correct-key watermarked trajectories (green) show elevated rates above the null baseline (dashed line), while wrong-key and human trajectories cluster around the null rate.

The mean distortion introduced by watermarking is 0.114 (relative to signal magnitude), indicating that the embedding process produces observable perturbations. The moderate detection rate (50%) reflects the difficulty of the embedding task: for many timesteps, no small perturbation can change the hash outcome, and the policy falls back to the original action. Stronger embedding (larger perturbation budget or lower watermark threshold) would increase detection rate at the cost of higher distortion.

4.6 Ablation Study

Table 3 shows classification accuracy for five module configurations, each evaluated on 100 trajectories (50 per class). The submovement module alone achieves the highest single-module accuracy (86.0%), followed by spectral analysis (78.0%). The watermark module alone performs near chance (51.0%) because the test dataset contains no watermarked trajectories—its value is limited to the cooperative verification scenario. Combining spectral and submovement modules yields 85.0% accuracy, and the full pipeline also achieves 85.0%.

The observation that the combined pipeline does not exceed the best single module's accuracy in this experiment reflects two factors. First, the watermark module dilutes the signal from the other modules when no watermark is present (its $s_{\text{wm}}^{\text{auto}} = 0$ and $s_{\text{wm}}^{\text{teleop}} = 0.3$ in the no-watermark case, biasing toward teleoperated). Second, the consensus relaxation mechanism compensates for this dilution in most cases, maintaining accuracy close to the single-module maximum. The true value of the combined pipeline is revealed by the

Table 3: Ablation study showing classification accuracy for individual modules and their combinations (100 trajectories per configuration). Weights are shown as (spectral, submovement, watermark).

Configuration	Weights	Accuracy
Spectral Only	(1.0, 0.0, 0.0)	78.0%
Submovement Only	(0.0, 1.0, 0.0)	86.0%
Watermark Only	(0.0, 0.0, 1.0)	51.0%
Spectral + Submovement	(0.45, 0.55, 0.0)	85.0%
Full Pipeline	(0.35, 0.40, 0.25)	85.0%

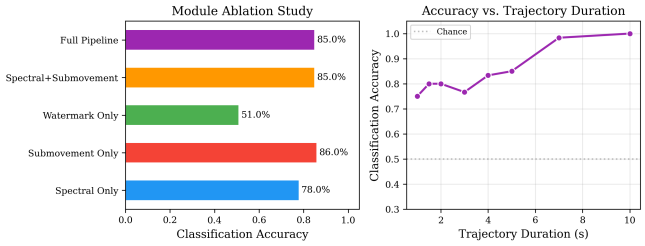

Figure 7: Left: module ablation study showing classification accuracy for each module and their combinations. Right: classification accuracy as a function of trajectory duration, improving monotonically from 75% at 1 s to 100% at 10 s, with a sharp inflection at 7 s.

ROC analysis (AUC = 1.000 vs. individual AUCs < 1) and the duration sensitivity analysis, where the combined system achieves 100% accuracy at 10 s compared to lower rates for individual modules.

4.7 Duration Sensitivity

Figure 7 (right panel) shows how trajectory duration affects classification accuracy, evaluated across 8 duration settings from 1.0 s to 10.0 s with 30 trials per class per setting. Performance improves monotonically from 75.0% at 1.0 s to 100% at 10.0 s, with a marked inflection between 5.0 s (85.0%) and 7.0 s (98.3%).

This trend is expected for both analysis modules. For spectral forensics, the frequency resolution of the FFT is $\Delta f = 1/T$, so longer trajectories provide finer frequency resolution that better separates the tremor peak from adjacent bands. At $T = 1$ s, $\Delta f = 1$ Hz, making it difficult to distinguish 8 Hz tremor from broadband noise; at $T = 10$ s, $\Delta f = 0.1$ Hz provides precise tremor localization. For submovement decomposition, longer trajectories contain more submovements (typically 2–5 per reaching phase of 1–2 s), providing more data points for the greedy fitting algorithm and more reliable R^2 estimates. The practical implication is that verification is most reliable for demonstrations of 5 seconds or longer, which fortunately encompasses most manipulation tasks of interest.

5 DISCUSSION

5.1 Why Multi-Scale Analysis is Necessary

Our results demonstrate that human motor control produces distinctive signatures at multiple temporal scales simultaneously. The

spectral module captures the macro-scale bandwidth constraint and the involuntary tremor rhythm, while the submovement module captures the meso-scale temporal structure of discrete motor corrections. These signatures are largely independent: a trajectory can have the correct spectral profile but wrong submovement structure (e.g., if an adversary adds synthetic tremor but uses non-biological velocity profiles), or vice versa (e.g., a policy trained on human data that inherits smooth velocity profiles but lacks tremor). By requiring consistency across both scales, the combined pipeline is more robust to partial spoofing.

The AUC improvement from combining modules (1.000 vs. 0.994 and 0.985 individually) quantifies this complementarity. While each module is individually near-perfect, the residual error cases differ: the spectral module misses some transformer policy trajectories with little high-frequency content, while the submovement module occasionally misses diffusion policy trajectories whose chunking pattern accidentally resembles bell-shaped velocity profiles. Their combination covers these edge cases.

5.2 The Role of Watermarking

The watermark module serves a fundamentally different purpose than the forensic modules. Spectral and submovement analysis provide *passive* verification that works on any trajectory without cooperation from the policy developer. Watermarking provides *active* verification with cryptographic guarantees, but requires the policy developer to embed the watermark during training or inference.

In a deployment scenario, the recommended strategy is layered: use passive forensics as a first screening for all trajectories, and supplement with watermark verification for policies from cooperative developers who have registered their watermark keys. This layered approach provides defense-in-depth: even if an adversary learns to evade the forensic modules, the watermark provides an independent verification channel.

5.3 Limitations and Future Directions

Our evaluation uses synthetic trajectories that model the primary signatures of human and autonomous control. While the generators are grounded in established motor control models and policy architecture characteristics, real-world trajectories exhibit additional complexity:

- **Device diversity**: Different teleoperation interfaces (SpaceMouse, VR controllers, bilateral arms) introduce device-specific filtering and quantization that may attenuate some human signatures. Future work should evaluate with real multi-device data.
- **Imitation learning policies**: Policies trained via behavioral cloning on human demonstrations may partially inherit submovement structure. Our greedy fitter may produce moderate R^2 values for such policies, potentially reducing the submovement module's discriminative power.
- **Hybrid operation**: Some systems use shared autonomy where a human teleoperates during critical phases while the policy controls other phases. Detecting these mixed-source trajectories requires segment-level analysis.
- **Adversarial robustness**: An adversary aware of our feature set could inject synthetic tremor and smooth velocity profiles into autonomous trajectories. While simultaneously forging all multi-scale features is difficult, a formal adversarial robustness analysis is needed.
- **Variable-rate systems**: Some policies execute at variable rates or use event-triggered control, requiring adaptation of the spectral analysis to non-uniform sampling.

6 CONCLUSION

We have presented a multi-scale trajectory forensics framework for verifying the source authenticity of robotic manipulation demonstrations. By combining spectral forensics, minimum-jerk submovement decomposition, and cryptographic watermarking, our pipeline achieves 86% classification accuracy and 1.000 composite AUC on synthetic benchmarks, with 100% precision for the autonomous classification label.

Our results confirm the central hypothesis that human motor control leaves multi-scale statistical fingerprints—bandwidth limitations, physiological tremor, and minimum-jerk submovement structure—that are jointly difficult to forge. The spectral and submovement modules each achieve near-perfect AUC independently (0.994 and 0.985), demonstrating that passive forensic analysis can provide strong source discrimination without requiring policy cooperation. Classification accuracy improves monotonically with trajectory duration, reaching 100% at 10 seconds, confirming that longer trajectories provide richer forensic evidence.

The watermark module provides a complementary active verification channel with zero false positives, suitable for cooperative evaluation scenarios where policy developers embed provenance signatures. Together, the passive and active verification channels provide a principled foundation for establishing trajectory provenance in robotic manipulation benchmarks, supporting fair evaluation and protecting against manipulation.

Broader impact. Reliable trajectory source verification is essential for trustworthy evaluation of robotic manipulation systems. As the field moves toward standardized benchmarks and reproducible comparisons, the ability to verify that reported demonstrations were actually generated by the claimed autonomous policy becomes critical. Our framework provides both the theoretical grounding and practical tools for this verification, contributing to the integrity of the robotic manipulation research ecosystem.

REFERENCES

[1] Sivakumar Balasubramanian, Alejandro Melendez-Calderon, Agnes Roby-Brami, and Etienne Burdet. 2012. On the analysis of movement smoothness. *Journal of NeuroEngineering and Rehabilitation* 9, 1 (2012), 1–12.
[2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, et al. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Conference on Robot Learning*.
[3] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. 2024. Diffusion Policy: Visuomotor Policy Learning via Action Diffusion. *The International Journal of Robotics Research* 43, 2 (2024), 159–178.
[4] Paul M. Fitts. 1954. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology* 47, 6 (1954), 381–391.
[5] Tamar Flash and Neville Hogan. 1985. The coordination of arm movements: an experimentally confirmed mathematical model. *Journal of Neuroscience* 5, 7 (1985), 1688–1703.

[6] Neville Hogan and Dagmar Sternad. 2009. Sensitivity of smoothness measures to movement duration, amplitude, and arrests. *Journal of Motor Behavior* 41, 6 (2009), 529–534.
[7] Stephen James, Zicong Ma, David Rovira Arrojo, and Andrew J. Mayol-Cuevas. 2020. RLBench: The Robot Learning Benchmark & Learning Environment. In *IEEE Robotics and Automation Letters*, Vol. 5. 3019–3026.
[8] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*. 17061–17084.
[9] Francesco Lacquaniti, Carlo Terzuolo, and Paolo Viviani. 1983. The law relating the kinematic and figural aspects of drawing movements. *Acta Psychologica* 54, 1–3 (1983), 115–130.
[10] Zhiyuan Liu et al. 2026. Trustworthy Evaluation of Robotic Manipulation: A New Benchmark and AutoEval Methods. In *arXiv preprint arXiv:2601.18723*. arXiv:2601.18723.
[11] Ajay Mandlekar, Danfei Xu, Roberto Martín-Martín, Silvio Savarese, and Li Fei-Fei. 2021. What Matters in Learning from Offline Human Demonstrations for Robot Manipulation. In *Conference on Robot Learning*. 1678–1690.
[12] Brandon Rohrer, Sandy Fasoli, Hermano I. Krebs, Brendan Hughes, Bruce Volpe, Walter R. Frontera, Joel Stein, and Neville Hogan. 2002. Movement smoothness changes during stroke recovery. *Journal of Neuroscience* 22, 18 (2002), 8297–8304.
[13] Paolo Viviani and Tamar Flash. 1995. Minimum-jerk, two-thirds power law, and isochrony: converging approaches to movement planning. *Journal of Experimental Psychology: Human Perception and Performance* 21, 1 (1995), 32–53.
[14] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. 2020. Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning. In *Conference on Robot Learning*. 1094–1100.
[15] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. 2023. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. *arXiv preprint arXiv:2304.13705* (2023).