Surviving the Unseen: Predictive Defense for Novel Multi-Turn Multimodal Attacks

Anonymous authors
Paper under double-blind review

Abstract

The expansion of Multimodal Large Language Models (MLLMs) and their integration into autonomous agentic workflows has introduced a non-stationary attack surface. Empirical observations indicate that adversaries employ progressive, cross-modal perturbations that evade turn-specific guardrails by distributing malicious intent across longitudinal conversational trajectories. Static defense mechanisms, constrained by the Markov property, evaluate inputs in isolation and fail to detect cumulative structural poisoning. To handle this limitation, this paper formulates safety verification as a dynamic survival prediction and trajectory dynamics problem. The Triple-tier Anomaly Defense (TRIAD) framework is proposed as a predictive model that maps multimodal and multi-turn conversational flow as a continuous trajectory. The framework integrates structural anomaly detection to monitor covariance shifts, a Ledoit-Wolf regularized Mahalanobis distance to monitor covariance shifts in high-dimensional spaces, and topological trajectory acceleration to differentiate benign creative exploration from continuous malicious drift. These kinematic and geometric features are integrated into a time-varying Cox Proportional Hazards model via a Bayesian Hidden Markov Model (HMM) feedback loop. Theoretical analysis demonstrates that the TRIAD framework provides a mathematically bounded expected time-to-failure under adversarial perturbations, ensuring that malicious acceleration diverges positively. This framework provides a computationally efficient, interpretable, and predictive safeguard for real-time agentic AI systems, establishing a rigorous foundation for continuous safety alignment without relying on empirical retraining.

1 Introduction

Generative Artificial Intelligence is transitioning from single-turn text systems into models capable of processing visual, auditory, and text information simultaneously over multi-turn interactions OpenAI et al. (2024); Team et al. (2025); Cui et al. (2026). While the integration of multi-modality maximizes reasoning capabilities and operational versatility, it introduces security vulnerabilities that are unseen in unimodal text environments where risk is measured via static snapshots. These vulnerabilities reveal a blind spot in legacy safety alignment systems, which often prioritize task instruction following over pre-programmed safety constraints (Bailey et al., 2024). Unlike small-scale applications where escalated issues can be manually reviewed case-by-case, production environments facing a massive user base require swift detection of unseen, nuanced adversarial attack patterns to prevent the unintentional generation of harmful content. Tackling these issues is a critical defense mechanism for maintaining continuous safeguard alignment.

When adversaries induce attention mismatches between text and image modalities, or progressively inject subtle instructions that appear harmless in individual turns, MLLMs showed their weakness to maintain their aligned safety policies Wei et al. (2023); Zou et al. (2023). This phenomenon functions as a cross-modal alignment tax. Static text-based guardrails and single-point binary classifiers deployed in production environments are constrained by the Markov Property; they evaluate the information of a specific input turn independently, lacking historical context (Albrethsen et al., 2026). Consequently, these defenses exhibit limitations in detecting structural poisoning caused by the long-term accumulation of subtle malicious contexts. If adversarial perturbations are distributed below the detection threshold of individual turns, the system ultimately loses control without recognizing the compounding risk (Russinovich et al., 2025; Liu et al., 2024).

The current empirical paradigm that relies heavily on large-scale defense datasets and supervised learning against specific, known attack signatures, this study investigated generative AI response in multimodal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and multi-turn environment as a dynamic survival prediction and trajectory dynamics problem over time. By treating the internal latent space of model as a multi-dimensional state space and modeling the continuous flow of conversation as a topological trajectory, a hybrid statistical framework is established to detect unknown zero-day attacks. The proposed defense mechanism is structured into a continuous analytical pipeline. 1) First, we employ Isolation Forest Liu et al. (2008) to identify structural isolation within the high-dimensional embedding space. 2) Second, it quantifies distributional drift via a robust distance metric that resolves singular matrix issues common in high-dimensional multimodal spaces Ledoit & Wolf (2004). 3) Third, it calculates trajectory acceleration via the second derivative of the phase space distance to statistically separate human creative exploration from adversarial evasion. 4) Finally, it predicts the time-to-failure hazard by integrating time-series indicators into a Survival Analysis model (Cox, 1972). This research provides mathematical validity that the proposed framework yields a bounded detection rate under unknown cross-modal poisoning. By adopting a computationally efficient anomaly detection approach, we ensure real-time feasibility for light immediate application in real-time agentic systems.

2 Related Work

Research on prompt jailbreaks and corresponding defenses in single-modality environments has been extensively conducted, but the advent of large-scale Vision-Language Models (VLMs) has expanded the attack surface Cui et al. (2026); Bailey et al. (2024). Optimized visual adversarial examples can neutralize the guardrails of safety-aligned language models, demonstrating that continuous, high-dimensional visual inputs bypass text-token-based filtering Wei et al. (2023). The security vulnerabilities of MLLMs under multi-turn prompting have become a central focus. Recent literature confirms vulnerabilities against automated multi-turn jailbreaks, such as Crescendo, where malicious intent is fragmented across interactions Russinovich et al. (2025); Liu et al. (2024). Also, recent frameworks like PolyJailbreak demonstrate that visual alignment introduces uneven safety constraints across modalities, leading to a phenomenon identified as multimodal safety asymmetry Wang et al. (2026). To exploit this, Foot-In-The-Door (FITD) methods leverage psychological manipulation to bypass safeguards across multiple turns by establishing benign context before escalating requests Weng et al. (2025). Advanced automated systems such as Mastermind employ a hierarchical multi-agent architecture with strategy-space fuzzing to autonomously discover and refine these multi-turn jailbreaks Songze Li et al. (2026). To counter these threats, the Multi-Turn Safety Alignment (MTSA) framework uses thought-guided attack learning and multi-turn reinforcement learning to improve the safety alignment of target models Singhania et al. (2025).

As Agentic AI systems transition to production, the threat landscape shifts toward indirect prompt injection, goal hijacking, and tool misuse. Unlike standard LLMs in read-only sandbox environments, Agentic AI systems possess read-write API access and persistent storage, escalating the impact of breaches from session-based misinformation to systemic compromise Chhabra et al. (2026). Evaluative frameworks confirm that existential and systemic safety remains a structural weakness, with current safety practices demonstrating partial alignment with emerging global standards but lacking rigorous implementation. Defensive strategies are shifting toward dynamic agent red teaming and test-time immunization, which aims to defend AI systems at the inference stage without requiring complete model retraining. To defend these multi-turn multimodal jailbreaks directly, systems use fragment-optimized MLLM defense mechanisms to systematically analyze and mitigate progressive attacks without requiring fine-tuning Das et al. (2026).

Out-of-Distribution (OOD) detection remains a core area of AI security for identifying unknown data patterns Lee et al. (2018); Ren et al. (2019). Previous frameworks utilized Gaussian Discriminant Analysis and Mahalanobis distance; however, directly applying this to the latent space of multimodal models encounters mathematical hurdles. To address the rank-deficiency and numerical instability inherent in high-dimensional multimodal manifoldsMueller & Hein (2025), we move beyond vanilla Mahalanobis metrics by incorporating robust shrinkage-based covariance estimation Ledoit & Wolf (2004). Recent advancements attempt to leverage the multimodal reasoning capabilities of MLLMs, synthesizing pseudo-OOD data from convex combinations of in-distribution data to enhance discrimination. Despite these advancements, robust statistical estimation integrating shrinkage techniques is required to mathematically handle the high dimensionality of multimodal embeddings.

To bypass strict Gaussian assumptions altogether, anomaly detection methodologies focusing on topological structure are utilized. While frameworks like Isolation Forest are traditionally deployed for large-scale network logs and API abuse detection, adapting them to monitor continuous embedding trajectories within multimodal interactions represents a paradigm shift (Liu et al., 2008; Pang et al., 2021; Ramprasath et al.,

2023). Recent paradigms build on this by combining embedding-based anomaly detection with LLM-driven validation frameworks. These systems dynamically address diverse anomaly types by pairing dimensionality reduction with Isolation Forests to filter out false positives typical of traditional machine learning detectors.

Rather than treating multi-turn interactions as independent events, we frame adversarial detection as a longitudinal survival problem. This approach is uniquely suited to capture the cumulative hazard of subtle semantic drifts—risks that are often invisible to memoryless, snapshot-based guardrails Cox (1972); Katzman et al. (2018). The application of survival analysis to LLM alignment has emerged as a method to model the robustness of models in extended multi-turn dialogues. Models such as Time-To-Inconsistency evaluate conversational failure via survival analysis, showing that gradual semantic drift is protective, while abrupt, prompt-to-prompt drift increases the hazard of policy violations Yubo Li et al. (2026). These studies analyze thousands of conversation turns to demonstrate that abrupt, prompt-to-prompt semantic drift increases the hazard of conversational failure. AFT models with model-drift interactions have achieved high discrimination and calibration in these contexts. However, these approaches rely heavily on semantic surface indicators within a single text modality, highlighting the need to apply survival analysis to the kinematic properties of multimodal latent spaces.

3 Problem Formulation and Framework Architecture

To overcome the limitations of static evaluation, the multimodal multi-turn interaction between a user and a Generative AI model is modeled as continuous state changes within a high-dimensional phase space. The interaction time-step is defined as discrete time $t \in \mathbb{N} = \{1, 2, ..., T\}$. At turn t, the user provides a text prompt $X_{\text{text}}^{(t)}$ and image data $X_{\text{img}}^{(t)}$, which the corresponding encoders map to latent representations $E_{\text{text}}^{(t)} \in \mathbb{R}^{d_{\text{text}}}$ and $E_{\text{img}}^{(t)} \in \mathbb{R}^{d_{\text{img}}}$. An integrated cross-modal semantic vector is generated using a concatenation operator:

$$V_{semantic}^{(t)} = E_{text}^{(t)} \oplus E_{img}^{(t)} \in \mathbb{R}^{d_{\text{semantic}}}. \tag{1}$$

To statistically separate the mechanical characteristics of an automated attacker from the cognitive reflection of a normal user, the Contextual Covariate Modulator (CCM) introduces a behavioral covariates vector $B^{(t)} \in \mathbb{R}^{d_{behav}}$. While known classical behavioral biometrics exploit static typing anomalies to detect localized automation Mehta et al. (2026), our framework treats these artifacts as a subset of a broader thermodynamic phenomenon: the low-entropy boundary condition of automated scripts. Because any automated adversarial framework—regardless of its underlying prompt architecture or unseen optimization target—must algorithmically dispatch tokens, it inherently collapses the temporal and physical variance of the interaction manifold. Conversely, benign creative exploration is characterized by high-entropy behavioral distributions, manifesting as stochastic temporal variations, irregular cognitive pauses, and non-linear trajectory accelerations across the interaction manifold. The final input state vector is defined as the concatenation of semantic and behavioral data:

$$V^{(t)} = V_{semantic}^{(t)} \oplus B^{(t)} \in \mathbb{R}^{D}, \tag{2}$$

where $D = d_{\text{semantic}} + d_{\text{behav}}$.

Traditional safety models evaluate a single vector $V^{(t)}$ to determine binary safety. However, under advanced multi-turn attacks, the maliciousness of initial turns approaches zero. The random variable T is defined as the number of turns at which the model first violates the aligned safety policy. The framework tracks the Survival Function $S(t) = \mathbb{P}(T > t)$, the probability that the dialogue survives without safety violations up to turn t. Correspondingly, the instantaneous rate at which the model loses control at the next turn t is formulated as the Hazard Function $h(t)$:

$$h(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \tag{3}$$

If malicious multi-turn perturbations are progressively injected, $S(t)$ drops exponentially alongside an increase in the hazard function $h(t)$.

To minimize latency while maximizing detection capability for real-time services, the TRIAD framework adopts a trigger-based cascade architecture combined with a Bayesian feedback loop.

Performing expensive matrix inversions across the entire phase space for every turn t is computationally inefficient. An Isolation Forest (iForest) ensemble is deployed as a primary unsupervised scout, possessing

Algorithm 1 Dynamic State-Space Inference Pipeline

Require: Sequence $\{X_{\text{text}}^{(t)}, X_{\text{img}}^{(t)}\}$, Behavioral data $B^{(t)}$, Params $\mu, \hat{\Sigma}_{\text{LW}}^{-1}$, thresholds $\tau_{\text{hazard}}, \alpha$

Initialize prior $\mathbb{P}(S_0) \leftarrow$ Safe
for each turn $t = 1, 2, \ldots$ **do**
  $V^{(t)} \leftarrow \text{Fusion}(X_{\text{text}}^{(t)}, X_{\text{img}}^{(t)}) \oplus B^{(t)}$
  $S_{\text{iso}}^{(t)} \leftarrow \text{iForest}(V^{(t)})$
  if $S_{\text{iso}}^{(t)} > \alpha$ **then**
    $D_M^{(t)} \leftarrow \sqrt{(V^{(t)} - \mu)^T \hat{\Sigma}_{\text{LW}}^{-1} (V^{(t)} - \mu)}$
    $a_t \leftarrow D_M^{(t)} - 2D_M^{(t-1)} + D_M^{(t-2)}$
    Update HMM Belief $\mathbb{P}(S_t \mid V^{(1:t)})$ using $D_M^{(t)}$ and a_t
    $h(t) \leftarrow h_0(t) \exp(\beta_1 D_M^{(t)} + \beta_2 S_{\text{iso}}^{(t)} + \gamma a_t)$
    if $h(t) > \tau_{\text{hazard}}$ **and** $a_t > 0$ **then**
      return ALERT: Imminent failure. Early-stop Session.
    end if
  end if
end for
return SAFE: Continue interaction.

a low time complexity of $\mathcal{O}(\log N)$. Given the expected path length $\mathbb{E}[h(V^{(t)})]$ required to isolate a vector through random feature space partitioning, where the expected path length in a binary search tree is defined by $c(n) = 2H(n-1) - \frac{2(n-1)}{n}$ and $H(i)$ is the harmonic number, the Anomaly Score $S_{iso}(V^{(t)})$ is calculated:

$$S_{iso}(V^{(t)}) = 2^{-\frac{\mathbb{E}[h(V^{(t)})]}{c(n)}}. \tag{4}$$

This score serves as a conditional trigger. If $S_{iso}(V^{(t)}) > \alpha$, the system allocates resources to execute precision computations.

When the system is triggered, the cross-modal covariance shift is measured. In high-dimensional spaces ($p \gg n$), calculating an accurate inverse covariance matrix Σ^{-1} is unstable due to rank deficiency. The sample covariance matrix is defined as $\Sigma = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$. To resolve the singular matrix problem, the Ledoit-Wolf Shrinkage estimator is employed. This method shrinks the empirical covariance matrix towards a well-conditioned target matrix, typically the identity matrix I, yielding:

$$\hat{\Sigma}_{LW} = (1 - \lambda)\Sigma + \lambda \frac{\text{Tr}(\Sigma)}{p} I, \tag{5}$$

where λ is optimized via Frobenius norm minimization. The robust Mahalanobis Distance is subsequently calculated:

$$D_M(V^{(t)}) = \sqrt{(V^{(t)} - \mu)^T \hat{\Sigma}_{LW}^{-1} (V^{(t)} - \mu)}. \tag{6}$$

Human dialogue naturally diverges temporarily to unfamiliar topics before recovering. Treating turns as disconnected independent events causes false permanent blocks. Within the CCM, the hidden state $S_t \in \{\text{Safe}, \text{Malicious}\}$ is mapped into a Hidden Markov Model (HMM). The posterior probability calculated in turn $t-1$ transfers as the prior probability for turn t, granting the system contextual inertia. This Bayesian state tracking ensures that the system evaluates current distances in the context of preceding trajectory stability.

To mathematically distinguish between a normal user's complex role-playing and an adversary's progressive attack, the topological Trajectory Acceleration is measured. Acceleration is approximated via the second-order derivative of the distance over discrete turns:

$$a_t = \frac{d^2}{dt^2} D_M(t) \approx D_M(t) - 2D_M(t-1) + D_M(t-2). \tag{7}$$

A benign user's creative search stabilizes over time within a new local manifold, causing the acceleration to converge toward zero or become negative. Conversely, a malicious Crescendo attack must continuously twist

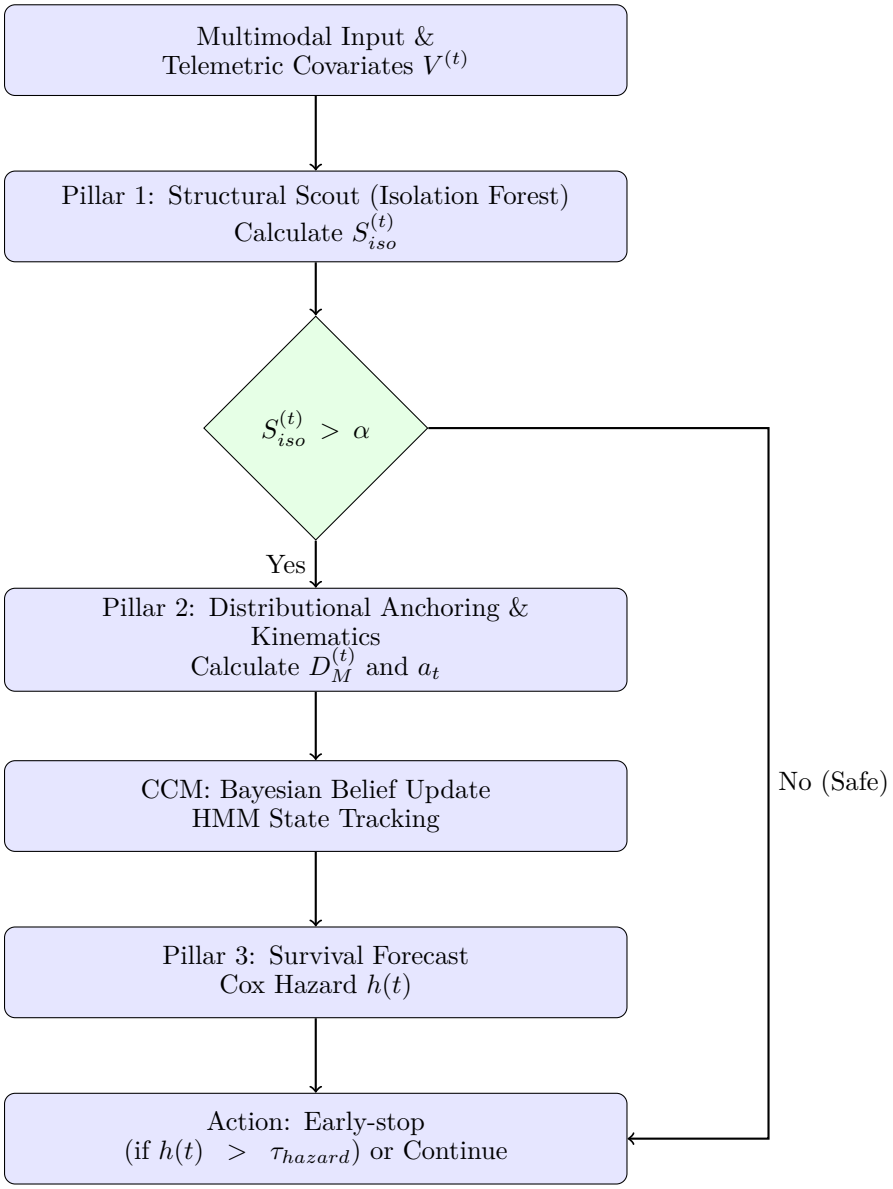

Figure 1: Operational Flow of the TRIAD Framework

and push the trajectory out of the manifold to bypass alignment guardrails, exhibiting an acceleration that strictly diverges in the positive direction ($a_t > 0$).

The extracted geometric isolation and kinematic acceleration indicators are processed into time-varying covariates $Z(t)$. These are injected into a Cox Proportional Hazards model to predict the instantaneous rate at which the model loses control:

$$h(t \mid Z(t)) = h_0(t) \exp\left(\sum_{i=1}^{k} \beta_i Z_i(t)\right). \tag{8}$$

The partial likelihood function maximized during training is defined as:

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(\beta^T Z_i(t_i))}{\sum_{j \in R(t_i)} \exp(\beta^T Z_j(t_i))}. \tag{9}$$

The term $h_0(t)$ represents the baseline hazard. Because the covariates $Z(t)$ and the positive acceleration metric a_t sit in the exponent, continuous adversarial drift creates a compound multiplier effect on the marginal hazard function $h(t)$.

4 Theoretical Analysis

This section rigorously demonstrates how the proposed state-space framework mathematically bounds the time of model failure under unknown multi-turn attacks, derived from the mathematical properties of the statistical shape modeling structure.

Theorem 1: Bounded Time-to-Failure under Cross-modal Poisoning. Assume an adversary applies a subtle malicious perturbation vector to the visual latent space to generate a manipulated image embedding $\tilde{E}_{img}^{(t)} = E_{img}^{(t)} + \delta$. This perturbation possesses an L_2 norm $||\delta||$ sufficient to induce malicious bias during the language model's multi-head cross-attention process, while maintaining a direction geometrically orthogonal to the existing text manifold. If the perturbation is continuously injected into the joint latent space, the time-series cumulative hazard function $H(t)$ estimated by the framework possesses a monotonically increasing property. The maximum number of conversation turns T_{max} the model can endure before its survival probability $S(t)$ drops below a specific safety threshold τ is mathematically bounded within a finite upper bound.

Proof: Let the covariance matrix calculated in the normal state's multivariate joint embedding space be represented using block matrix partitioning. Substituting the joint state vector of a new discrete time-step with the perturbation vector added as $\tilde{V}^{(t)} = V^{(t)} + [0, \delta]^T$ into the Ledoit-Wolf distance metric projection yields:

$$D_M(\tilde{V}^{(t)})^2 = (\tilde{V}^{(t)} - \mu)^T \hat{\Sigma}_{LW}^{-1} (\tilde{V}^{(t)} - \mu). \tag{10}$$

Expanding the equation by the distributive law yields:

$$D_M(\tilde{V}^{(t)})^2 = D_M(V^{(t)})^2 + 2[0, \delta]^T \hat{\Sigma}_{LW}^{-1} (V^{(t)} - \mu) + \delta_{sub}^T (\hat{\Sigma}_{LW}^{-1})_{I,I} \delta_{sub}. \tag{11}$$

By the properties of matrix regularization, the Ledoit-Wolf estimate applied with the shrinkage scalar $\lambda > 0$ is guaranteed to be a strictly positive definite matrix with a stabilized condition number. By definition, for any non-zero vector ($\delta \neq 0$), the quadratic form $\delta^T \hat{\Sigma}_{LW}^{-1} \delta$ yields a strictly positive scalar. The squared distance increment induced by the perturbation $\Delta D_M(t)^2$ secures a lower bound constant $c||\delta||^2 > 0$ (where c is the minimum eigenvalue λ_{min} of the inverse matrix).

Approximating the continuous Cox survival function into a series of discrete time-steps results in:

$$S(t) \approx \exp\left(-\sum_{k=1}^{t} h_0(k) \exp(\beta_1 D_M(k) + \beta_2 S_{iso}(k)) \right). \tag{12}$$

As malicious perturbations intervene at every turn, a strictly positive exponential increment is added to the covariate component. As $t \to \infty$, according to the divergence test of infinite series, the summation diverges toward infinity. Consequently, the dependent survival probability function $S(t) = \exp(-\infty)$ rapidly meets an exponential convergence to 0. It is proven that the point T_{max} satisfying the minimum safe survival tolerance $\tau > 0$ must finitely exist.

Theorem 2: Positive Divergence of Adversarial Acceleration. Assume normal creative exploration converges to a local manifold μ_{local} representing a topological transition. Malicious Crescendo attacks follow continuous drift Brownian motion, constantly diffusing to evade detection thresholds without settling. Under continuous adversarial drift required to bypass alignment guardrails, the second-order derivative (trajectory acceleration) $a_t = \frac{d^2}{dt^2} D_M(t)$ remains strictly positive, distinguishing it mathematically from the asymptotic stabilization of benign creative shifts.

Proof: Let $V(t)$ be the state position function parameterized by discrete time t. Under a benign creative search scenario, $\lim_{t\to\infty} V(t) = \mu_{local}$ holds true. By the chain rule of calculus, the first derivative of the Mahalanobis distance scalar function $D_M(V(t))$ is derived as $\nabla D_M \cdot V'(t)$. As $V(t)$ approaches the new local settlement point, the magnitude of the velocity vector $V'(t)$ generated at each turn asymptotically decays toward 0. Hence, the trajectory acceleration function $\frac{d^2}{dt^2} D_M(t)$ converges to 0 or becomes negative.

Conversely, for an adversarial attacker to bypass the aligned filter network, they must continuously inject a forced displacement ΔV above a constant threshold at every time-step ($||V'(t)|| \geq c > 0$ for a valid constant c). Since such a trajectory continuously causes friction against the structural cohesion of the model's existing covariance matrix, the magnitude of the directional gradient component ∇D_M shows a monotonically increasing form. In the second derivative expansion equation:

$$\frac{d^2}{dt^2} D_M(t) = V'(t)^T H_{D_M} V'(t) + \nabla D_M \cdot V''(t). \tag{13}$$

(where H_{D_M} is the Hessian matrix), the sum of the two terms fails to turn negative due to the continuously added drift tension. Therefore, malicious acceleration never converges to 0 and always exceeds a positive lower bound.

5 Practical Deployment and Experimental Validation Guidelines

Deploying state-space statistical estimation models in production enterprise environments necessitates mitigating inherent structural constraints and operational edge cases. Instead of anchoring validation to volatile, model-specific empirical benchmarks, this section establishes a rigorous *Experimental Protocol* and evaluation framework to systematically stress-test multi-turn agentic defenses against unknown adversarial drift.

5.1 Architectural Mitigations for Operational Constraints

Manifold Heterogeneity via GMM: The core mathematical derivation initially assumes that the multimodal interaction manifold approximates a single Multivariate Gaussian Distribution. However, in diverse production deployments, the joint latent space exhibits multi-peak distributions representing disparate topical clusters. Utilizing a single global covariance matrix causes the distance calculation to artificially expand when a user naturally switches conversational domains, inducing a high rate of false positives. To mitigate this, implementations should adopt a Gaussian Mixture Model (GMM) framework. During runtime, the system should probabilistically identify the local cluster K most appropriate for the current semantic context, calculating the regional deviation distance using the regional mean μ_K and the regional shrunken covariance matrix $\hat{\Sigma}_{K,LW}^{-1}$.

Temporal Inertia via AFT Models: The Cox-based survival model is predicated on the Proportional Hazards Assumption, making it vulnerable to sudden shock attacks. An adversary could orchestrate a high-latency sequence to accumulate safe inertia within the system, followed by the sudden injection of a substantial jailbreak payload utilizing the vast context window limit at a late turn. To prevent historical inertia from suppressing urgent risk surges, production pipelines must integrate an ensemble approach employing Accelerated Failure Time (AFT) models. Utilizing Weibull distributions, AFT models directly shrink the baseline survival time based on the immediate impact magnitude of the most recent turn.

Stochastic Auditing for Sub-Threshold Evasion: The trigger-based cascade architecture conditionally performs the intensive Mahalanobis computation only when the iForest score exceeds the threshold α. An adversary adopting a boiling frog micro-perturbation approach could meticulously maintain the anomaly score below α across consecutive turns. Deployments must implement a stochastic checkpoint scheduling mechanism as an asynchronous background process. This mechanism forcefully executes a full covariance precision inspection periodically and completely independent of the iForest score, ensuring cumulative micro-adversarial adjustments are quantified.

5.2 Standardized Adversarial Simulation Protocol

To validate the efficacy of the TRIAD framework under unknown zero-day attacks without relying on known attack signatures, we propose a standardized cross-modal context poisoning simulation protocol:

1. **Baseline Parameterization:** Extract normal multimodal interaction embeddings from standard conversational datasets to establish reference covariance structures and baseline isolation dimensions.
2. **Orthogonal Drift Injection:** Simulate a stealthy multi-turn attack by gradually injecting an orthogonal noise vector δ into the visual or textual latent spaces over successive interaction intervals. This bounds individual turn perturbations below standard single-point classification thresholds.

3. **Kinematic Profile Evaluation:** Monitor the second-order derivative (trajectory acceleration a_t) across the sequence. A valid defense framework must demonstrate that while baseline snapshot detectors fail across extended turns, the continuous accumulation of drift tension reflects a strictly positive acceleration, forcing the survival probability $S(t)$ to exponentially converge to 0 before session completion.

5.3 Computational Complexity Boundaries

For real-time agentic AI deployments where latency is a critical constraint, evaluation protocols must verify computational feasibility boundaries. In offline initialization or periodic calibration phases, calculating the robust inverse covariance matrix and constructing the Isolation Forest requires $O(Nd^2 + N \log N)$ time. Crucially, during online real-time inference, evaluating the incoming telemetric vector requires only $O(d^2)$ matrix operations and $O(\log N)$ tree traversal time. This mathematically guarantees deterministic, low-latency execution independent of the length of the semantic interaction history.

6 Conclusion

As generative artificial intelligence evolves toward complex, multi-turn Agentic systems, the mechanisms governing alignment and safety must achieve operational parity through advanced, non-linear geometric tracking. Traditional classification models are structurally inadequate against the progressive cross-modal contamination inherent in multi-turn adversarial interactions. By reformulating safety alignment from a reactive classification challenge to a predictive, time-to-failure survival process, the TRIAD framework resolves fundamental blind spots in current architectures. The integration of unsupervised Isolation Forests, Ledoit-Wolf regularized Mahalanobis metrics, and topological kinematics yields an active state-space immune system capable of functioning independent of exhaustive attack signature repositories. Tracking the second-order derivative of semantic divergence provides a robust statistical demarcation between benign creative exploration and continuous adversarial drift, offering a highly efficient mechanism for securing next-generation artificial intelligence.

Despite its theoretical guarantees, practical deployment of the TRIAD framework requires careful handling of operational edge cases, such as manifold heterogeneity, temporal inertia from sudden shock attacks, and sub-threshold evasion techniques. Incorporating structural extensions like Gaussian Mixture Models (GMM) and Accelerated Failure Time (AFT) frameworks helps preserve accuracy during natural conversational domain shifts while neutralizing high-latency adversarial strategies. Transitioning from static telemetric thresholds to dynamic, closed-loop calibration loops ensures resilience without triggering excessive false positives during complex, benign interactions.

Consequently, this methodology marks a decisive shift from static, reactive prompt-filtering to dynamic, predictive survival modeling in AI alignment, establishing the statistical foundations required to operationalize compliance and safety metrics for high-risk autonomous infrastructure. As the industry transitions toward fully autonomous multi-turn systems handling sensitive societal frameworks, defining a mathematical lower bound for adversarial drift detection enables continuous risk auditing in clinical and financial decision-making pipelines, thereby accelerating the safe and responsible integration of generative agents into public infrastructure.

Ultimately, while this paper establishes the mathematical foundation for predictive statistical defense, real-world implementations must evolve into a complex, interdependent mesh network of feedback loops to prevent sophisticated adversaries from reverse-engineering and exploiting the defense mechanism itself.

7 Ethics Statement

This paper focuses on the theoretical foundations of securing Multi-modal Large Language Models (MLLMs) against multi-turn adversarial attacks. As a theoretical defense framework, this research does not involve human subjects, proprietary data collection, or the release of a commercial product. The primary ethical consideration lies in the analyzing the trajectory dynamics of multi-turn attacks requires detailing the mechanics of how static guardrails might be bypassed. However, the explicit intent and contribution of the proposed TRIAD framework are defensive for such scenario. By establishing mathematically bounded detection mechanisms for structural poisoning, this research aims to advance AI trust and safety so that it

can support the responsible deployment of autonomous agentic systems reliable when there is unseen attack types happen.

8 Reproducibility Statement

To ensure the reproducibility of our theoretical claims, all mathematical assumptions and proofs are made transparent and logically self-contained within the manuscript. The complete derivations for the bounded time-to-failure under cross-modal poisoning (Theorem 1) and the positive divergence of adversarial acceleration (Theorem 2) are detailed in Section 4 and 5. Furthermore, to bridge our theoretical findings with practical verification, Section 5.2 explicitly outlines a standardized adversarial simulation protocol. This section details the baseline parameterization and orthogonal drift injection methodologies required for independent researchers to simulate and verify the framework's kinematic properties using standard conversational datasets.

9 AI Use Statement

In this work, we used generative AI tools solely to assist with proofreading and editing the manuscript to improve readability and grammatical correctness. We did not use generative AI tools to develop theoretical models, formulate mathematical claims, derive proofs, or generate synthetic data.

References

Justin Albrethsen, Yash Datta graves, Kunal Kumar, and Sharath Rajasekar. DeepContext: Stateful real-time detection of multi-turn adversarial intent drift in LLMs, 2026. URL https://arxiv.org/abs/2602.16935. arXiv preprint arXiv:2602.16935.

Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: Adversarial images can control generative models at runtime, 2024. URL https://arxiv.org/abs/2309.00236. arXiv preprint arXiv:2309.00236.

Anshuman Chhabra, Shrestha Datta, Shahriar Kabir Nahin, and Prasant Mohapatra. Agentic AI security: Threats, defenses, evaluation, and open challenges, 2026. URL https://arxiv.org/abs/2510.23883. arXiv preprint arXiv:2510.23883.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Kaiyuan Cui, Yige Li, Yutao Wu, Xingjun Ma, Sarah Erfani, Christopher Leckie, and Hanxun Huang. Toward universal and transferable jailbreak attacks on vision-language models, 2026. URL https://arxiv.org/abs/2602.01025. arXiv preprint arXiv:2602.01025.

Badhan Chandra Das, Md Tasnim Jawad, Joaquin Molto, M. Hadi Amini, and Yanzhao Wu. Multi-turn jailbreaking attack in multi-modal large language models, 2026. URL https://arxiv.org/abs/2601.05339. arXiv preprint arXiv:2601.05339.

Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, feb 2018. ISSN 1471-2288. doi: 10.1186/s12874-018-0482-1. URL https://doi.org/10.1186/s12874-018-0482-1.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. ISSN 0047-259X. doi: 10.1016/S0047-259X(03)00096-4. URL https://www.sciencedirect.com/science/article/pii/S0047259X03000964.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks, 2018. URL https://arxiv.org/abs/1807.03888. arXiv preprint arXiv:1807.03888.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*, pp. 413–422, 2008. doi: 10.1109/ICDM.2008.17.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. AutoDAN: Generating stealthy jailbreak prompts on aligned large language models, 2024. URL https://arxiv.org/abs/2310.04451. arXiv preprint arXiv:2310.04451.

Atharva Mehta, Rajesh Kumar, Aman Singla, Kartik Bisht, Yaman Kumar Singla, and Rajiv Ratn Shah. Detecting LLM-assisted academic dishonesty using keystroke dynamics, 2026. URL https://arxiv.org/abs/2511.12468. arXiv preprint arXiv:2511.12468.

Maximilian Mueller and Matthias Hein. Mahalanobis++: Improving OOD detection via feature normalization, 2025. URL https://arxiv.org/abs/2505.18032. arXiv preprint arXiv:2505.18032.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, et al. GPT-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774. arXiv preprint arXiv:2303.08774.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Comput. Surv.*, 54(2), mar 2021. doi: 10.1145/3439950. URL https://doi.org/10.1145/3439950.

J. Ramprasath, S. Ramakrishnan, V. Tharani, R. Sushmitha, and D. Arunima. Cloud service anomaly traffic detection using Random Forest. In Shailesh Tiwari, Munesh C. Trivedi, Mohan L. Kolhe, and Brajesh Kumar Singh (eds.), *Advances in Data and Information Sciences*, pp. 269–279, Singapore, 2023. Springer Nature. ISBN 978-981-19-5292-0.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A. DePristo, Joshua V. Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection, 2019. URL https://arxiv.org/abs/1906.02845. arXiv preprint arXiv:1906.02845.

Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo multi-turn LLM jailbreak attack, 2025. URL https://arxiv.org/abs/2404.01833. arXiv preprint arXiv:2404.01833.

Abhishek Singhania, Christophe Dupuy, Shivam Mangale, and Amani Namboori. Multi-lingual multi-turn automated red teaming for LLMs, 2025. URL https://arxiv.org/abs/2504.03174. arXiv preprint arXiv:2504.03174.

Songze Li, Ruishi He, Xiaojun Jia, Jun Wang, and Zhihui Fu. Knowledge-driven multi-turn jailbreaking on large language models, 2026. URL https://arxiv.org/abs/2601.05445. arXiv preprint arXiv:2601.05445.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, et al. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805. arXiv preprint arXiv:2312.11805.

Xinkai Wang, Beibei Li, Zerui Shao, Ao Liu, Guangquan Xu, and Shouling Ji. PolyJailbreak: Cross-modal jailbreaking attacks on black-box multimodal LLMs, 2026. URL https://arxiv.org/abs/2510.17277. arXiv preprint arXiv:2510.17277.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does LLM safety training fail?, 2023. URL https://arxiv.org/abs/2307.02483. arXiv preprint arXiv:2307.02483.

Zixuan Weng, Xiaolong Jin, Jinyuan Jia Regel, and Xiangyu Zhang. Foot-in-the-door: A multi-turn jailbreak for LLMs, 2025. URL https://arxiv.org/abs/2502.19820. arXiv preprint arXiv:2502.19820.

Yubo Li, Ramayya Krishnan, and Rema Padman. Time-to-inconsistency: A survival analysis of large language model robustness to adversarial attacks, 2026. URL https://arxiv.org/abs/2510.02712. arXiv preprint arXiv:2510.02712.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043. arXiv preprint arXiv:2307.15043.