Evaluating Vulnerabilities in Vision-Language Models: Impact of Behavior-Induced Interference

Yuwei Chen, Shiyong Chu*
Aviation Industry Development Research Center of China
Aviation Key Laboratory of Science and Technology on Unmanned Blue Force
No.14 Xiao Guan Dong Li, Chaoyang District, Beijing, China
catcornic@gmail.com, csy319ldl@163.com

Abstract

Vision-language models (VLMs) are increasingly recognized as central components for multimodal scene understanding and decision-support in complex autonomous systems. While existing research on VLM robustness has predominantly focused on digital-domain perturbations such as pixel-level noise or adversarial prompt injection, adversarial influence in dynamic environments can also manifest through structured changes in observable behavior. In this work, we investigate the susceptibility of VLMs to behavior-induced perturbations within system dynamics, rather than the sensor data stream. Leveraging a high-fidelity simulation environment, we design a series of behavior-level variations, including motion pattern changes, altitude adjustments, and emission reconfiguration, that preserve platform identity and task semantics while altering observable action patterns. To quantify decision instability, we introduce two metrics: Ranking Drift, which measures overall shifts in priority ranking, and Priority Inversion Rate, which captures cases where the highest-priority entity is erroneously deprioritized. Evaluations across multiple state-of-the-art multimodal foundation models demonstrate that even semantically invariant behavioral variations can induce significant Ranking Drift and frequent priority inversion. Our results reveal a critical and underexplored vulnerability: behavioral salience alone can systematically bias prioritization in multimodal reasoning pipelines. These findings highlight the urgent need for robustness evaluations that consider behavior-level dynamics when deploying VLMs in safety-critical autonomous systems.

1. Introduction

The emergence of VLMs has fundamentally transformed the capacity for artificial intelligence to execute complex multimodal reasoning and interpret dynamic visual scenes.

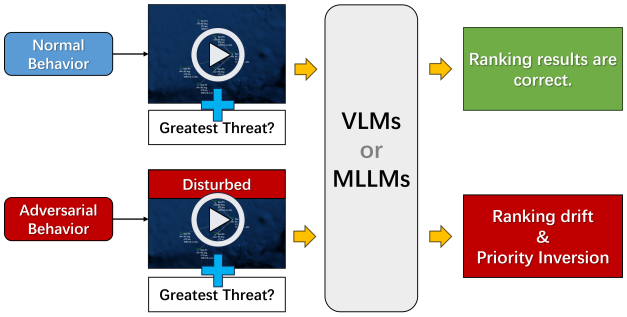

Figure 1. Illustration of behavior-induced vulnerability in vision-language models. Under normal behavior, the model correctly identifies the highest-threat target. When adversarial behavioral variations are introduced without changing platform capability or mission context, the model may produce ranking drift or priority inversion.

By integrating visual encoders with natural language processing, these models facilitate sophisticated visual question answering and decision-support tasks[1, 14, 32]. Consequently, there is growing interest in deploying VLMs within safety-critical sectors, such as autonomous surveillance and unmanned aerial vehicle (UAV) operations, where they assist in real-time monitoring, situational awareness, and threat assessment[12, 17, 19, 20, 31].

However, the rapid adoption of these models has outpaced our understanding of their robustness in adversarial or non-idealized settings. Current research into multimodal robustness predominantly targets *input-level manipulations*, such as adversarial patches[16, 27, 36], digital image noise[24, 25], or prompt-based attacks[11, 26, 29]. While these studies highlight significant vulnerabilities, they generally assume an adversary possesses the ability to directly manipulate the digital sensory data or the textual instructions ingested by the model.

In real-world dynamic environments, decision systems may be influenced not only by changes in sensor inputs,

but also by variations in observable behavior. This extends the scope of interference to the *physical and semantic domains*, where motion patterns, altitude profiles, or emission states change over time while the underlying platform capabilities and task objectives remain invariant. We characterize this phenomenon as **behavior-induced interference**, which aligns with the broader taxonomy of *semantics-preserving physical attacks*. Unlike digital perturbations, this form of interference occupies a unique position in robustness research: it exploits the model's *inherent heuristic biases*—where the system naturally over-prioritizes visual salience—and serves as a *stealthy adversarial attack vector* that misleads decision logic through legitimate maneuvers without requiring access to digital bitstreams.

This observation raises a fundamental question: can behavior-level variations bias priority judgments in vision-language models, even when the underlying task-relevant structure remains unchanged? Addressing this question is essential for the reliable deployment of autonomous systems that must achieve *semantic-level robustness* in dynamic scenes and allocate attention across multiple observed entities.

We investigate this phenomenon by constructing a controlled UAV observation scenario within the Command Modern Operations (CMO) simulation environment. The scenario involves multiple platforms with identical physical capabilities, evolving under stable conditions. We then introduce a systematic set of behavior-induced perturbations, including formation dispersion, altitude variation, radar emission switching, electronic reconfiguration, and localized maneuver patterns. These perturbations are designed to preserve task-relevant semantics while altering observable behavioral salience. We evaluate how several state-of-the-art VLMs adjust their ranking outputs in response to these variations.

To quantify the resulting decision shifts, we introduce two metrics: **Ranking Drift**, which measures the magnitude of change in model ranking outputs relative to a baseline, and **Priority Inversion Rate**, which tracks cases in which the ground-truth highest-priority entity is deprioritized. Our experiments demonstrate that even modest behavioral variations can induce measurable Ranking Drift and frequent priority inversion, revealing a fundamental misalignment between behavioral salience and the model's *reasoning-layer* priority structure.

The primary contributions of this work are as follows:

- We define behavior-level perturbations as a distinct perspective for analyzing VLM robustness, identifying them as a source of *semantic-level vulnerability* that bridges inherent cognitive bias and strategic physical-domain attacks.
- We construct a controlled UAV observation setting in which behavior-level variations can be introduced while preserving task-relevant semantics, enabling systematic analysis of decision instability under semantically invariant conditions.
- We demonstrate that behavioral signals alone can induce Ranking Drift and Priority Inversion across diverse multimodal foundation models, revealing that current architectures often conflate *perceptual salience* with *operational priority*.

2. Related Work

2.1. Vision-Language Models for Multimodal Scene Understanding

Recent advancements in VLMs and multimodal large language models (MLLMs) have established a robust capacity for grounding visual evidence in linguistic outputs. These developments have enabled transformative applications in autonomous scene understanding [1, 10, 14], visual question answering [7, 9, 18], and complex instruction-following across image and video domains [4, 5]. Beyond passive perception, research is increasingly positioning VLMs as decision-support components within interactive and dynamic environments. In these settings, models must synthesize temporally evolving observations and generate strategic responses while adhering to operational constraints [22, 23]. This shift toward real-time application necessitates a rigorous analysis of robustness in operationally realistic environments where decisions rely on complex multimodal evidence rather than curated benchmarks.

2.2. Adversarial Robustness and Security of VLMs

A substantial body of literature demonstrates that VLMs are susceptible to adversarial manipulation, particularly through perturbations within the visual input space[2, 3, 13, 26, 28, 29, 34, 35]. Recent studies have introduced scalable attack pipelines and black-box strategies to generate adversarial examples [8, 34, 35], illustrating that visually grounded reasoning can be derailed without direct modification of model parameters. Notable research directions include large-scale adversarial generation and transfer-based black-box attacks, as well as the implementation of naturalistic patches that maintain visual plausibility while inducing systematic model failure.

Simultaneously, the security of multimodal systems has been scrutinized through the lens of safety and jailbreak attacks. These works reveal that even safety-aligned MLLMs can be coerced into producing undesired outputs via meticulously crafted multimodal prompts or adversarial visual triggers [3, 30, 33]. While defense strategies such as adversarial tuning [21, 35, 36] and consistency-based enhancement methods [6, 13, 15] have emerged, they primarily address vulnerabilities within the input space. These existing

frameworks typically assume that an adversary has the requisite access to alter the digital sensory inputs or the textual instructions delivered to the model.

2.3. Behavior-Induced Decision Vulnerability

Despite the progress in understanding adversarial attacks on VLMs, a critical dimension remains largely unexplored: adversarial influence through behavior-level manipulation in dynamic environments. In many high-stakes operational settings, adversaries may be unable to tamper with digital sensor data but can effectively modify their observable behavior through legitimate actions. These actions include variations in motion patterns, the restructuring of formations, adjustments to altitude profiles, or the modulation of electronic emissions.

These behavior-level variations alter the multimodal evidence perceived by a decision system while preserving the underlying identity and operational capabilities of the platform. Our work complements the existing literature by investigating perturbations applied at the level of system dynamics rather than at the pixel or prompt level. By examining how semantically valid behavioral transformations induce decision drift in priority assessment, we provide empirical evidence that behavioral salience can mislead multimodal decision outputs, even when the objective operational context remains invariant.

3. Methodology

3.1. VLM-Based Decision Formulation

We formalize a dynamic observation environment wherein a vision-language model serves as the primary decision-support engine. The comprehensive system state at time t is represented as

$$\mathbf{x}_t \in \mathcal{X}, \tag{1}$$

where $\mathcal{X}$ denotes the manifold of task-relevant physical and operational variables, including platform kinematics, emission statuses, and spatial configurations.

Multimodal observations are generated through a perception mapping function g, such that

$$\mathbf{o}_t = g(\mathbf{x}_t), \tag{2}$$

where $g : \mathcal{X} \to \mathcal{O}$ transforms the latent system state into the multimodal observation space $\mathcal{O}$. In our framework, $\mathcal{O}$ encompasses both high-resolution visual renderings of the environment and structured symbolic data typically available to a monitoring system.

Given a semantic task instruction $q \in \mathcal{Q}$ (e.g., "Prioritize the detected entities by importance"), the VLM synthesizes the information to produce a decision output

$$\mathbf{y}_t = F_\theta(\mathbf{o}_{1:t}, q), \tag{3}$$

where F_θ represents the parameterized foundation model and $\mathbf{o}_{1:t}$ denotes the temporal sequence of observations accumulated up to time t. The output $\mathbf{y}_t \in \mathcal{Y}$ resides within a decision space that, in the context of this study, is operationalized as an ordered ranking of detected entities based on perceived priority.

Equations (1)–(3) provide a unified mathematical abstraction for VLM-driven decision-making in dynamic observation scenarios. This formulation allows us to systematically analyze how structured behavioral perturbations, introduced at the state level $\mathcal{X}$, propagate through the perception mapping $g(\cdot)$ to influence the final decision output $\mathbf{y}_t$.

3.2. Behavior-Induced Adversarial Operator

We formalize behavior-induced interference as a structured perturbation applied within the state space rather than directly to the sensory manifold. Let the state trajectory over a temporal horizon be denoted as $\mathbf{x}_{1:T} \in \mathcal{X}^T$. We define a behavior-induced adversarial operator $\mathcal{A}$ such that

$$\tilde{\mathbf{x}}_{1:T} = \mathcal{A}(\mathbf{x}_{1:T}; \alpha), \tag{4}$$

where $\mathcal{A}$ represents a parameterized behavioral transformation and α determines the intensity or duration of the perturbation.

A fundamental requirement of this operator is the preservation of task-relevant invariants, ensuring that the perturbation does not alter the underlying task context. We define a semantic mapping function $\Phi : \mathcal{X}^T \to \mathcal{Z}$, which projects the high-dimensional state trajectory into a latent semantic space $\mathcal{Z}$ representing core task-relevant attributes. To address the mapping rules of Φ, we decompose $\mathcal{Z}$ into a set of observable invariants $\mathcal{Z} = \{z_{id}, z_{pos}, z_{goal}\}$, where z_{id} represents the platform's physical identity (e.g., UAV type and capabilities), z_{pos} denotes the global formation centroid, and z_{goal} signifies the intended mission trajectory.

The invariance condition is satisfied when:

$$\|\Phi(\tilde{\mathbf{x}}_{1:T}) - \Phi(\mathbf{x}_{1:T})\| < \epsilon, \tag{5}$$

where ϵ is a task-specific threshold. Quantitatively, this constraint is enforced through specific mapping rules:

- **Identity Invariance:** $\text{Type}(\tilde{\mathbf{x}}) = \text{Type}(\mathbf{x})$, ensuring no digital-domain modification to the asset's visual features.
- **Spatial Centroid Invariance:** $\sum_{j=1}^{N} \tilde{p}_t^j = \sum_{j=1}^{N} p_t^j$, where p_t^j is the position of the j-th agent, ensuring the aggregate threat location remains constant.
- **Kinematic Feasibility:** $\dot{\tilde{\mathbf{x}}} \in \mathcal{V}_{feasible}$, ensuring the perturbed behavior remains within the platform's legitimate operational envelope.

This formalized constraint ensures that while the observable manifestations (e.g., dispersion or altitude) change, the underlying task-relevant significance remains constant. Integrating this with the perception and decision framework,

the perturbed decision output is:

$$\tilde{\mathbf{y}}_t = F_\theta(g(\tilde{\mathbf{x}}_{1:t}), q). \tag{6}$$

By comparing $\tilde{\mathbf{y}}_t$ with the baseline $\mathbf{y}_t$ under the constraint in Eq. (5), we can effectively decouple decision-level shifts induced by behavioral salience from those necessitated by actual changes in the operational environment.

3.3. Decision Drift under Behavioral Perturbation

Building on the framework established in Eqs. (4)–(6), we define the decision drift induced by a behavioral perturbation $\mathcal{A}$ at time t as

$$\Delta\mathbf{y}_t(\mathcal{A}) = \mathcal{D}(\mathbf{y}_t, \tilde{\mathbf{y}}_t), \tag{7}$$

where $\mathcal{D} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ represents a distance metric or discrepancy function that quantifies the divergence between the baseline and perturbed decision outputs.

By substituting the perception and decision mappings into Eq. (7), the drift is explicitly characterized as a functional of the behavioral perturbation operator:

$$\Delta\mathbf{y}_t(\mathcal{A}) = \mathcal{D}\Big(F_\theta(g(\mathbf{x}_{1:t}), q), F_\theta(g(\mathcal{A}(\mathbf{x}_{1:t})), q)\Big). \tag{8}$$

Crucially, under the invariance constraint defined in Eq. (5), the semantic ground truth remains constant across both decision states. Consequently, any non-zero value of $\Delta\mathbf{y}_t(\mathcal{A})$ signifies a decision-making instability that arises solely from behavioral re-expression rather than from an evolution in the underlying task-relevant semantics. This formulation enables the systematic isolation and measurement of behavioral salience as a distinct vulnerability in multimodal reasoning.

4. Experiments and Results

4.1. Model Capability Validation

Prior to evaluating the impact of behavior-induced perturbations, we conduct a preliminary capability validation to ensure that each VLM can accurately ground fundamental visual elements within the simulation environment. This calibration step is essential to ensure that any observed decision drift in subsequent stages is a product of reasoning instability rather than a failure in basic perceptual recognition.

We evaluate a diverse suite of state-of-the-art multimodal foundation models, including Qwen-3.5-Plus, GPT-5.2, Gemini-3, ChatGLM-5, Ernie-5.0, Claude-4.6-Opus, and Hunyuan-Thinking-1.5. To establish a controlled baseline, we utilize a video sequence depicting six platforms maintaining uniform heading and velocity without any tactical maneuvers. Each model is provided with a grounding prompt: "This video is a recorded session from a simulation. Please identify the total number of entities displayed in the scene."

Our validation results show that Qwen-3.5-Plus and ChatGLM-5 demonstrate reliable performance under direct video stream inputs. For these models, subsequent experiments proceed using continuous video sequences as the observation history $\mathbf{o}_{1:t}$. The remaining models exhibit inconsistent grounding performance with raw video; consequently, we provide these models with a sequence of temporally ordered keyframes extracted from critical decision points. This modality-aware adjustment satisfies the requirements of the perception mapping $g(\cdot)$ while preserving the behavioral semantics of the scenario, thereby ensuring that the foundation for higher-level decision-making remains stable across all tested architectures.

This validation phase is not intended for benchmark comparison. Rather, it serves to establish a standardized perceptual baseline, ensuring that any subsequent instances of Ranking Drift or Priority Inversion can be rigorously attributed to behavior-induced perturbations at the state level $\mathcal{X}$ rather than primitive recognition errors.

4.2. Experimental Setup

To instantiate the decision framework formulated in Section 3.1, we construct a controlled dynamic observation scenario within the CMO simulation environment. This setup is designed to isolate the influence of behavior-induced perturbations on model outputs while maintaining strict consistency in task-relevant semantics.

The baseline configuration consists of six platforms maintaining uniform heading and velocity under stable formation. In this reference phase, no additional motion changes, altitude fluctuations, or emission-state transitions occur. The rendered interface view is presented to the VLM together with a standardized natural language instruction q, requesting a priority ranking of the observed entities. The resulting output $\mathbf{y}_t$ serves as the reference decision under behaviorally neutral conditions.

Building upon this baseline, we instantiate the perturbation operator $\mathcal{A}$ defined in Eq. (4) by applying structured transformations to specific components of the state variable $\mathbf{x}_t$ from Eq. (1). Every perturbation is designed to satisfy the invariance constraint $\Phi(\tilde{\mathbf{x}}_{1:T}) = \Phi(\mathbf{x}_{1:T})$ in Eq. (5), ensuring that platform identity and task-relevant properties remain unchanged.

Specifically, we implement four classes of behavioral transformations within a bounded temporal window $[t_0, t_1]$:

Formation Dispersion. This spatial restructuring modifies the relative positioning of platforms without altering the collective trajectory or velocity magnitude:

$$\tilde{\mathbf{p}}_t^j = \mathbf{p}_t^j + \delta_j, \tag{9}$$

where the displacement vectors $\{\delta_j\}$ are constrained by $\sum_{j=1}^{N} \delta_j = \mathbf{0}$. This ensures that the formation centroid and global motion vector remain constant.

Altitude Modification. Vertical profile perturbations adjust the altitude h while preserving horizontal velocity components:

$$\tilde{h}_t^j = h_t^j - r(t - t_0), \tag{10}$$

where r denotes the rate of ascent or descent.

Emission-State Reconfiguration. Observable signaling patterns are modified through state transitions:

$$\tilde{e}_t^j = \mathcal{E}(e_t^j), \quad \tilde{m}_t^j = \mathcal{M}(m_t^j; \beta), \tag{11}$$

where $\mathcal{E}$ and $\mathcal{M}$ denote operators for emission-state switching and localized configuration changes, respectively.

Localized Motion Patterns. Trajectory restructuring introduces bounded orbital motion around a stable reference point $\mathbf{c}$:

$$\tilde{\mathbf{p}}_t^j = \mathbf{c} + \begin{bmatrix} R\cos\omega(t - t_0) \\ R\sin\omega(t - t_0) \\ 0 \end{bmatrix}, \tag{12}$$

where R is the orbital radius and ω is the angular frequency.

The perturbed observation sequences $\tilde{\mathbf{o}}_{1:t}$ are provided to each VLM via its validated modality (video or sequential keyframes). Throughout all trials, the task instruction remains fixed. By leveraging the observation mapping in Eq. (2) and the decision function in Eq. (3), we systematically evaluate how these state-level variations propagate into the decision-level drift characterized by Eq. (8).

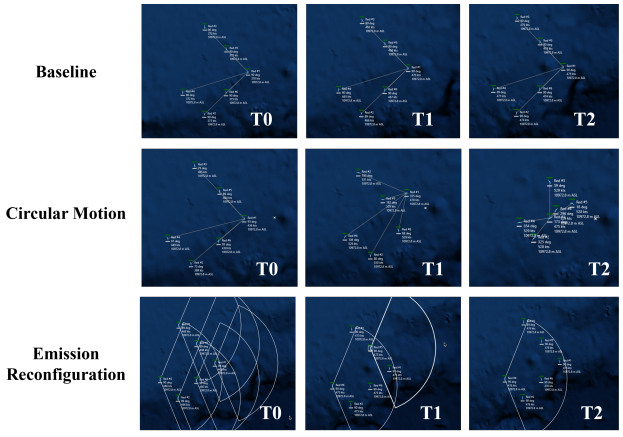

Figure 2. Representative behavioral perturbation sequences (T0–T2). Top row: baseline scenario featuring a stable formation with invariant motion and emission states. Middle row: localized motion perturbation introducing bounded circular motion patterns. Bottom row: emission-state reconfiguration in which observable signaling states are modulated while platform identity and task-relevant context remain unchanged. These sequences illustrate the decoupling of observable behavioral salience from underlying task-relevant semantics.

Figure 2 provides a visual representation of the transition from baseline sequences to perturbed observation histories used to evaluate VLM decision stability.

4.3. Quantitative Evaluation

We evaluate behavior-induced vulnerability by quantifying the divergence between decision outputs under baseline and perturbed states. For each model and perturbation type, we first establish a baseline ranking $\pi^{(0)}$ from the stable scenario described in Section 4.2. Subsequently, we apply the behavior-induced operator $\mathcal{A}$ to obtain the perturbed ranking $\pi^{(a)}$ under identical task instructions q.

4.3.1. Evaluation Metrics

In accordance with the decision drift functional established in Eq. (8), the *Ranking Drift* (RD) is calculated as the Spearman footrule distance between the baseline and perturbed rankings:

$$RD = \sum_{i=1}^{N} \left| \pi^{(0)}(i) - \pi^{(a)}(i) \right|. \tag{13}$$

For evaluations spanning multiple time steps within a perturbation window, RD is reported as the temporal average across that window.

To measure critical decision failure, let i^* denote the ground-truth primary entity defined by the scenario parameters. The inversion indicator $\mathbb{I}$ for a single trial is defined as:

$$\mathbb{I} = \mathbf{1}\{\pi^{(a)}(i^*) > 1\}. \tag{14}$$

Aggregating across M trials or time steps, the *Priority Inversion Rate* (PIR) is expressed as:

$$PIR = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}\{\pi_m^{(a)}(i^*) > 1\}. \tag{15}$$

Unless otherwise specified, all reported RD and PIR values represent averages across the four perturbation categories and their respective evaluation windows.

4.3.2. Ground-truth Specification

In our constructed simulation, the entity designated as "Primary Entity #1" is established as the ground-truth primary priority (i^*). This assignment is based on the heuristic that the leading entity closest to the point of interest is most likely to trigger an event, whereas trailing entities occupy secondary roles. This configuration allows us to evaluate whether behavior-induced perturbations compel the model to deprioritize i^* despite unchanged platform capabilities and task context.

Table 1. Behavior-induced vulnerability metrics across different input modalities.

Model	Video + Text		Keyframes + Text	
	RD ↓	PIR ↓	RD ↓	PIR ↓
Qwen-3.5-Plus	7.2	0.0	–	–
ChatGLM-5	10.4	0.6	–	–
GPT-5.2	–	–	6.4	0.4
Gemini-3	–	–	3.6	0.2
Ernie-5.0	–	–	6.4	0.4
Claude-4.6-Opus	–	–	6.4	1.0
Hunyuan-Thinking-1.5	–	–	10.4	0.6
Average	**8.8**	**0.3**	**6.64**	**0.52**

4.3.3. Evaluation Protocol

For each VLM, four perturbation categories are instantiated as described in Section 4.2: formation dispersion, altitude modification, emission switching, electronic reconfiguration, and localized circular motion. Each perturbation generates an evaluation trial consisting of the perturbed observation sequence and the corresponding model-generated priority ranking.

A consistent task instruction is maintained across all trials to ensure decision-making uniformity. The baseline ranking $\pi^{(0)}$ is derived from the stable phase, while the perturbed ranking $\pi^{(a)}$ is produced under the perturbation operator $\mathcal{A}$. For each trial, we compute the RD and the inversion indicator, yielding aggregated metrics by averaging across the four perturbation categories.

4.3.4. Results

Table 1 summarizes the aggregated vulnerability metrics across different input modalities. Models demonstrating reliable video grounding are evaluated under the *Video + Text* condition, while those with unstable video perception are evaluated using temporally ordered keyframes (*Keyframes + Text*) to ensure perceptual stability.

Across all evaluated models, behavior-induced perturbations consistently produce measurable Ranking Drift despite the invariance of platform identities and task semantics. This indicates that subtle behavioral variations can significantly alter the decision prioritization logic within current multimodal reasoning pipelines.

Among video-capable models, ChatGLM-5 exhibits the most significant instability ($RD = 10.4$) alongside a non-zero inversion rate ($PIR = 0.6$), demonstrating that multiple behavioral shifts can trigger the deprioritization of the primary entity. Conversely, Qwen-3.5-Plus maintains a PIR of 0.0 despite moderate Ranking Drift, suggesting a more robust internal hierarchy.

In the keyframe-based evaluation, Gemini-3 displays the highest stability, yielding the lowest RD (3.6) and PIR (0.2). However, other models remain highly sensitive to behavioral cues. Most notably, Claude-4.6-Opus produces a PIR of 1.0, indicating that the primary entity was misprioritized in every perturbation trial. Similarly, Hunyuan-Thinking-1.5 shows substantial Ranking Drift ($RD = 10.4$) and a high inversion rate ($PIR = 0.6$), confirming that observable behavioral signals can override task-relevant priority in diverse foundation models.

4.4. Behavioral Vulnerability Analysis

The quantitative results in Section 4.3 indicate that behavior-induced perturbations consistently destabilize priority ranking in state-of-the-art VLMs. Although underlying platform identities, capabilities, and task contexts remain invariant, subtle variations in observable behavior elicit measurable Ranking Drift across diverse architectures. This section explores the potential mechanisms underlying this observed vulnerability.

4.4.1. Sensitivity to Behavioral Salience

A primary explanation for the observed instability is that VLMs exhibit an acute sensitivity to observable behavioral cues during multimodal reasoning. Signals such as motion patterns, altitude fluctuations, and changes in emission states are frequently misinterpreted as indicators of significant operational changes. Consequently, entities exhibiting higher visual or behavioral salience may receive disproportionate attention, overshadowing more stable but contextually important variables.

Crucially, in our controlled simulation scenarios, these behavioral signals are decoupled from actual changes in platform capabilities. Nevertheless, the experimental data demonstrate that such variations are sufficient to reconfigure the model's internal ranking system. This suggests that behavioral salience functions as a potent implicit cue, potentially overriding structural logic within multimodal decision pipelines.

4.4.2. Heuristic Reasoning Bias

The observed decision drift likely reflects heuristic reasoning strategies learned during large-scale pretraining. VLMs are typically trained on extensive visual-textual corpora where specific behavioral patterns—such as rapid movement or active signal emission—frequently co-occur with task-relevant contexts or descriptions of unusual activity.

When applied to the simulation environment, these learned associations manifest as an inductive bias: the model prioritizes entities that appear behaviorally anomalous rather than those that pose the greatest risk based on spatial proximity or task-relevant factors. This indicates that priority assessment in current foundation models is governed more by perceived behavioral cues than by objective positioning or temporal factors.

4.4.3. Influence of Input Modality

The results further suggest that the representation modality modulates the stability of priority ranking. Models processing temporally ordered keyframes generally exhibit lower Ranking Drift compared with video-based evaluations. A plausible explanation is that discrete frames effectively filter out transient motion noise, forcing the model to rely on stable spatial relationships.

In contrast, continuous video input accentuates dynamic motion patterns, which may amplify the perceived salience of behavioral variations. This divergence implies that the temporal resolution of visual information interacts with behavioral cues to determine decision stability, with higher temporal fidelity potentially increasing susceptibility to behavioral interference.

4.4.4. Perturbation Sensitivity Analysis

To isolate which behavioral factors exert the most significant influence on VLM decision-making, we analyze the sensitivity across the four perturbation categories defined in Section 4.2.

Formation Dispersion. This spatial restructuring modifies the relative positioning of entities without altering the collective trajectory or velocity magnitude. Consequently, most models maintain a stable ranking under this perturbation, as the global motion vectors and task configurations remain visually consistent.

Altitude Modification. Vertical motion introduces moderate Ranking Drift by altering the perceived profile of entities. While the model may interpret this as a change in posture, the horizontal trajectories and relative configurations remain unchanged, limiting the resulting prioritization shifts.

Emission-State Reconfiguration. Changes in emission states or electronic signaling introduce significant decision variance. These signals are visually and semantically distinctive, often misinterpreted by the model as indicators of shifting intent or operational status, leading to fluctuations in the ranking of entities. This category is the most frequent driver of Priority Inversion.

Localized Motion Patterns. Entities deviating from a linear formation trajectory to perform orbital motions appear behaviorally anomalous. This deviation attracts a higher weight in the model's reasoning process, often causing the entity to be erroneously prioritized due to its increased visual activity.

4.4.5. Discussion on Perturbation Intensity and Gradients

While this work primarily establishes the existence of behavior-induced vulnerabilities using semantic-preserving perturbations of fixed intensity, the relationship between perturbation magnitude (e.g., the rate of altitude change or the radius of orbital motion) and decision instability warrants further discussion.

Preliminary observations suggest that the Ranking Drift is not necessarily linear with respect to the intensity of behavioral shifts. Instead, VLMs appear to exhibit "threshold-based" sensitivity, where a behavioral change becomes an adversarial cue only after exceeding a certain level of visual salience. Below this threshold, the model's spatial reasoning remains dominant; above it, the behavioral heuristic takes precedence. This non-linear mapping between physical intensity and decision drift suggests that VLM vulnerability is deeply tied to the model's internal attention allocation rather than simple geometric transformations. A systematic exploration of these intensity gradients and their impact on PIR will be a core focus of our future research.

In summary, behavioral signals associated with electronic activity and anomalous motion patterns exert the strongest influence on model outputs. Conversely, perturbations that preserve stable formation dynamics result in significantly lower decision drift, highlighting that visual activity is often conflated with operational priority in current multimodal architectures.

5. Discussion

Our experiments demonstrate that behavior-induced perturbations can significantly destabilize priority ranking in VLMs, even when the underlying platform identities, capabilities, and task context remain strictly invariant.

5.1. Behavioral Salience vs. Task-Relevant Priority Grounding

The results suggest that priority ranking in contemporary VLMs is heavily influenced by behavioral salience. Entities exhibiting visually distinctive actions, such as motion patterns, emission shifts, or trajectory deviations, frequently attract higher attention weights during multimodal reasoning. Within the model's reasoning pipeline, these signals are interpreted as implicit indicators of heightened activity, intent, or anomalous status.

However, in our controlled observation setting, these behavioral cues are intentionally decoupled from task-relevant context. While all entities maintain identical task objectives and capabilities, and the ground-truth priority hierarchy is governed by task-specific factors, the models consistently recalibrate their prioritization in response to observable behavioral shifts. This divergence indicates that priority grounding in VLMs may be disproportionately influenced by perceived behavioral changes rather than stable task-relevant factors like spatial relationships or task structure. Consequently, behavioral salience functions as a latent bias that can override objective priority assessment.

5.2. Behavioral Manipulation as a Dynamic Attack Surface

These findings reveal that behavioral signals themselves constitute a sophisticated manipulation channel within autonomous decision pipelines, representing a distinct form of vulnerability compared to traditional digital-domain perturbations. While conventional adversarial research primarily focuses on input-level corruptions (e.g., pixel-level noise or prompt injection targeting the low-level sensory manifold), the behavior-induced interference investigated here operates entirely at the semantic and reasoning level. In this paradigm, an entity modifies its observable actions without altering its physical capabilities or direct sensor data representation, exploiting the model's internal heuristic reasoning rather than its signal-processing integrity.

Such modifications are particularly impactful in dynamic environments where changes in motion patterns, emission control, or trajectory adjustments are legitimate behaviors. This legitimacy makes behavioral manipulation significantly harder to detect using conventional security frameworks, which are typically designed to identify statistical anomalies or data-stream corruption. An adversary could strategically deploy behavior patterns that appear anomalous or highly salient to divert the attention of automated systems from higher-priority but visually stable entities. The severity of this vulnerability lies in its ability to bypass robustness at the perception layer; our results suggest that a model may remain resilient to digital noise yet fail catastrophically when confronted with strategic deception through physically valid behavior.

5.3. Mitigation Strategies and Implications

Our observations have critical implications for the deployment of autonomous systems that rely on multimodal foundation models for situational reasoning. To address the identified vulnerabilities, we propose several potential mitigation strategies. First, *logic-augmented prompting* can be employed to incorporate explicit physical invariants, including kinematic history or distance-to-target, into the model's reasoning chain, forcing a recalibration based on objective variables rather than visual salience. Second, *diversity-driven behavioral augmentation* during the fine-tuning stage could expose models to a wider envelope of anomalous but semantically invariant maneuvers, improving their immunity to heuristic biases. Finally, implementing *cross-modal consensus filtering* can provide a safety layer where VLM-based rankings are cross-referenced with traditional rule-based threat assessments to detect and flag attention-ranking anomalies.

In safety-critical applications such as surveillance and resource allocation, behavioral deception could otherwise lead to significant attention misallocation. Entities performing high-salience behaviors could divert resources away from higher-priority risks that maintain a low-profile behavioral signature. In real-time operations, such misprioritization could compromise monitoring accuracy and overall situational awareness, necessitating the development of the aforementioned reasoning-layer defenses to distinguish between superficial behavioral salience and actual operational risk.

6. Conclusion

This study systematically characterizes the susceptibility of VLMs to behavior-induced perturbations within the context of dynamic wargame decision-making. By leveraging a high-fidelity simulation environment, we instantiated a series of structured behavioral modifications—ranging from formation dispersion and altitude fluctuations to electronic reconfiguration and localized maneuvers—that preserve platform identity and mission semantics while altering observable action patterns.

Our evaluation across a diverse suite of multimodal foundation models reveals that behavioral variations can profoundly destabilize threat prioritization. Despite a constant operational threat structure, the tested VLMs exhibited significant Ranking Drift and, in several instances, catastrophic Priority Inversion. These findings provide empirical evidence that multimodal reasoning in contemporary architectures is often anchored in transient behavioral salience rather than in stable, task-relevant operational factors such as engagement geometry or kinetic intent.

This research identifies a critical vulnerability in autonomous decision pipelines: adversarial influence can manifest not only through the digital manipulation of sensory inputs but also through the strategic re-expression of observable behaviors. Because these behavioral modifications constitute legitimate tactical actions in physical environments, this form of influence is exceptionally difficult to detect or mitigate through conventional cybersecurity or input-level defense frameworks.

Moving forward, the development of robust autonomous systems will necessitate the integration of temporal consistency constraints, cross-modal threat grounding, and uncertainty-aware decision calibration. Decoupling behavioral salience from objective mission semantics remains an essential challenge for the safe and reliable deployment of multimodal foundation models in high-stakes surveillance and autonomous decision-support applications.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 2

[2] Hongge Chen, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, and Cho-Jui Hsieh. Attacking visual language grounding with adversarial examples: A case study on neural image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2587–2597, 2018. 2

[3] Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24625–24634, 2024. 2

[4] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023. 2

[5] Shengyuan Ding, Shenxi Wu, Xiangyu Zhao, Yuhang Zang, Haodong Duan, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Mm-ifengine: Towards multimodal instruction following. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1099–1109, 2025. 2

[6] Xixi Ga, Wenjie Liu, Tongyu Zhu, Shan Kou, Meishen Liu, and Yue Hu. Evaluating robustness and diversity in visual question answering using multimodal large language models. *OSF Preprints*, 2024. 2

[7] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10867–10877, 2023. 2

[8] Qi Guo, Shanmin Pang, Xiaojun Jia, Yang Liu, and Qing Guo. Efficient generation of targeted and transferable adversarial examples for vision-language models via diffusion models. *IEEE Transactions on Information Forensics and Security*, 20:1333–1348, 2024. 2

[9] Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review. *Frontiers in artificial intelligence*, 7:1430984, 2024. 2

[10] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 2

[11] Md Zarif Hossain and Ahmed Imteaj. Securing vision-language models with a robust encoder against jailbreak and adversarial attacks. In *2024 IEEE International Conference on Big Data (BigData)*, pages 6250–6259. IEEE, 2024. 1

[12] Yibiao Hu, You Zhou, Zhengqiang Zhu, Xi Yang, Han Zhang, Kun Bian, and Hong Han. Llvm-drone: A synergistic framework integrating large language models and vision models for visual tasks in unmanned aerial vehicles. *Knowledge-Based Systems*, page 114190, 2025. 1

[13] Chengze Jiang, Zhuangzhuang Wang, Minjing Dong, and Jie Gui. Survey of adversarial robustness in multimodal large language models. *arXiv preprint arXiv:2503.13962*, 2025. 2

[14] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *European conference on computer vision*, pages 105–124. Springer, 2022. 1, 2

[15] Zaid Khan and Yun Fu. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 10854–10863, 2024. 2

[16] Dehong Kong, Siyuan Liang, Xiaopeng Zhu, Yuansheng Zhong, and Wenqi Ren. Patch is enough: naturalistic adversarial patch against vision-language pre-training models. *Visual Intelligence*, 2(1):33, 2024. 1

[17] Maroš Krupáš, L'ubomír Urblík, and Iveta Zolotová. Multimodal ai for uav: Vision–language models in human–machine collaboration. *Electronics*, 14(17):3548, 2025. 1

[18] Jusung Lee, Sungguk Cha, Younghyun Lee, and Cheoljong Yang. Visual question answering instruction: Unlocking multimodal large language model to domain-specific visual multitasks. *arXiv preprint arXiv:2402.08360*, 2024. 2

[19] Ye Li, Li Yang, Meifang Yang, Fei Yan, Tonghua Liu, Chensi Guo, and Rufeng Chen. Navblip: a visual-language model for enhancing unmanned aerial vehicles navigation and object detection. *Frontiers in Neurorobotics*, 18:1513354, 2025. 1

[20] Pouya Parsa, Keya Li, Kara M Kockelman, and Seongjin Choi. Video-based vehicle surveillance in the wild: License plate, make, and model recognition with self reflective vision-language models. *arXiv preprint arXiv:2508.01387*, 2025. 1

[21] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 21527–21536, 2024. 2

[22] Zuojin Tang, Bin Hu, Chenyang Zhao, De Ma, Gang Pan, and Bin Liu. Vlascd: A visual language action model for simultaneous chatting and decision making. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 9223–9243, 2025. 2

[23] Zhipeng Tang, Sha Zhang, Jiajun Deng, Chenjie Wang, Guoliang You, Yuting Huang, Xinrui Lin, and Yanyong Zhang. Vlmplanner: Integrating visual language models with motion planning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 5040–5049, 2025. 2

[24] Muhammad Usama, Syeda Aishah Asim, Syed Bilal Ali, Syed Talal Wasim, and Umair Bin Mansoor. Analysing the robustness of vision-language-models to common corruptions. *arXiv preprint arXiv:2504.13690*, 2025. 1

[25] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024. 1

[26] Taowen Wang, Zheng Fang, Haochen Xue, Chong Zhang, Mingyu Jin, Wujiang Xu, Dong Shu, Shanchieh Yang, Zhenting Wang, and Dongfang Liu. Large vision-language

model security: A survey. In *International Conference on Frontiers in Cyber Security*, pages 3–22. Springer, 2024. 1, 2

[27] Yubo Wang, Chaohu Liu, Yanqiu Qu, Haoyu Cao, Deqiang Jiang, and Linli Xu. Break the visual perception: Adversarial attacks targeting encoded visual tokens of large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 1072–1081, 2024. 1

[28] Peng Xie, Yequan Bie, Jianda Mao, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. Chain of attack: On the robustness of vision-language models against transfer-based adversarial attacks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14679–14689, 2025. 2

[29] Mang Ye, Xuankun Rong, Wenke Huang, Bo Du, Nenghai Yu, and Dacheng Tao. A survey of safety on large vision-language models: Attacks, defenses and evaluations. *arXiv preprint arXiv:2502.14881*, 2025. 1, 2

[30] Gulsum Yigit and Mehmet Fatih Amasyali. From text to multimodal: a survey of adversarial example generation in question answering systems. *Knowledge and Information Systems*, 66(12):7165–7204, 2024. 2

[31] Tongtong Yuan, Xuange Zhang, Bo Liu, Kun Liu, Jian Jin, and Zhenzhen Jiao. Surveillance video-and-language understanding: from small to large multimodal models. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(1):300–314, 2024. 1

[32] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024. 1

[33] Shaobo Zhang, Wenli Chen, Xiong Li, Qin Liu, and Guojun Wang. Apbam: Adversarial perturbation-driven backdoor attack in multimodal learning. *Information Sciences*, 700: 121847, 2025. 2

[34] Tianyuan Zhang, Lu Wang, Xinwei Zhang, Yitong Zhang, Boyi Jia, Siyuan Liang, Shengshan Hu, Qiang Fu, Aishan Liu, and Xianglong Liu. Visual adversarial attack on vision-language models for autonomous driving. *arXiv preprint arXiv:2411.18275*, 2024. 2

[35] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36:54111–54138, 2023. 2

[36] Wanqi Zhou, Shuanghao Bai, Danilo P Mandic, Qibin Zhao, and Badong Chen. Revisiting the adversarial robustness of vision language models: a multimodal perspective. *arXiv preprint arXiv:2404.19287*, 2024. 1, 2