Hierarchical Attacks for Multi-Modal Multi-Agent Reasoning

Hao Zhou* Tiru Wu* Yan Jiang* Wanqi Zhou Junxing Hu Ai Han†
JD.com
hanai5@jd.com

Abstract

Multi-modal multi-agent systems (MM-MAS) have gained increasing attention for their capacity to enable complex reasoning and coordination across diverse modalities. As these systems continue to expand in scale and functionality, investigating their potential vulnerabilities has become increasingly important. However, existing studies on adversarial attacks in multi-agent systems primarily focus on isolated agents or unimodal settings, leaving the vulnerabilities of MM-MAS largely underexplored. To bridge this gap, we introduce HAM^3, a Hierarchical Attack framework for multi-modal multi-agent systems that decomposes attacks into three interconnected layers. Specifically, at the perception layer, HAM^3 mounts attacks by perturbing visual inputs, textual inputs, and their fused visual–textual representations. At the communication layer, it performs communication-level attacks that corrupt message content and interaction topology, such as manipulating shared context or communication links to distort collective information flow. At the reasoning layer, it conducts reasoning-level attacks that interfere with each agent's cognitive pipeline, biasing reasoning trajectories and ultimately compromising final decisions. We evaluate HAM^3 on the GQA benchmark through multi-agent systems built on distinct reasoning paradigms including ReAct, Plan-and-Solve, and Reflexion. Experiments demonstrate that our framework achieves an Attack Success Rate of up to 78.3%, with reasoning-layer attacks being the most effective. More than half of the successful attacks lead multiple agents to produce consistent errors. These findings offer valuable insights for building more robust and interpretable multi-agent intelligence.

1. Introduction

Recent progress [16, 21, 22, 40] in multi-modal and multi-agent learning has made collaborative perception and decision-making increasingly important, driving deployment across diverse domains such as social interaction [31], embodied control [39], and autonomous driving [19]. As these collaborative systems expand in scale and interconnectivity, studying their adversarial vulnerabilities is increasingly crucial for ensuring reliable and resilient multi-agent intelligence [10].

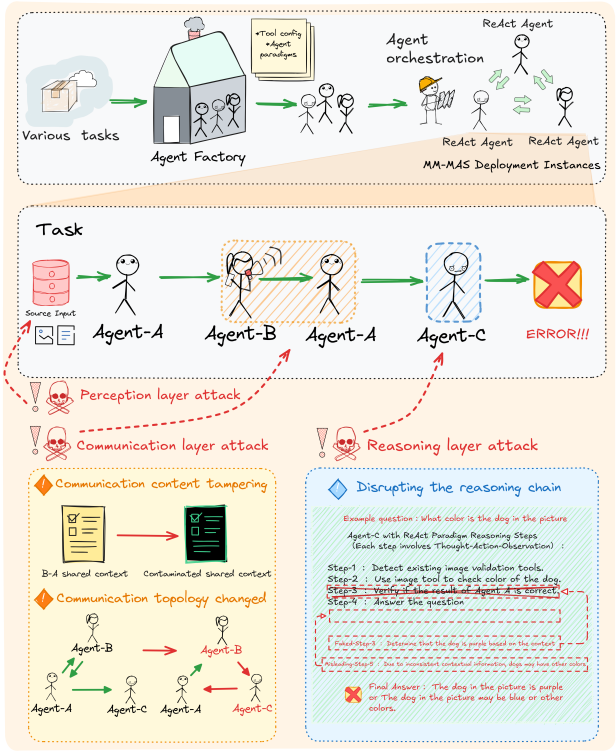

Figure 1. Hierarchical Attacks on a multimodal multi-agent system, illustrating the three layers of attack: perception, communication, and reasoning. The diagram depicts the lifecycle of a multi-agent system, highlighting the attack manifestations at each layer and providing examples of how these attacks affect the system's functionality.

Prior work on adversarial vulnerabilities has primarily centered on single-agent settings, where attackers manipulate observations, prompts, or memory to bias an individual agent's reasoning [3, 4]. Recent multi-agent attack studies largely extend single-agent adversarial principles to multi-

*Equal contribution.
†Corresponding author.

agent settings, primarily by perturbing agent-specific messages or manipulating shared functional interfaces to influence individual decision-making [9, 23]. Although such approaches reveal weaknesses in inter-agent message exchange and coordination mechanisms, they remain confined to content-level manipulations. Consequently, they fall short of examining structural vulnerabilities rooted in communication topology or collective reasoning dynamics that arise uniquely in multi-agent systems. In parallel, research on multimodal adversarial attacks largely targets model-level perception, such as typographic, compositional, or logic-based visual prompts that jailbreak or mislead vision-language models [6, 30, 47, 49], rather than attacking the agentic decision-making pipeline. As a result, adversarial robustness of multimodal LLM-based agents, especially under multi-agent collaboration, remains substantially underexplored.

To address these limitations, as illustrated in Figure 1, we introduce HAM^3, a unified adversarial framework that characterizes how perturbations propagate across the perception, communication, and reasoning layers of multimodal multi-agent systems. The **perception layer** models adversarial manipulations to visual, textual, or other multimodal inputs that influence all agents at the entry point. The **communication layer** captures disruptions in inter-agent information flow, including message tampering, chain blocking, and agent impersonation, which alter both message content and interaction topology. The **reasoning layer** formalizes interference within each agent's internal inference process, where attacks either directly modify intermediate reasoning steps or indirectly bias the contextual signals that guide downstream inference. Collectively, these components provide a structured view of how localized perturbations can cascade through the multi-agent workflow and compromise the final collective decision.

Our contributions are threefold:

- We conduct the first systematic investigation of adversarial robustness in *multimodal multi-agent* systems and introduce a multimodal agent attack benchmark, which will be publicly available.
- We propose HAM^3, a unified adversarial framework that decomposes perturbation effects across perception, communication, and reasoning layers, characterizing how localized attacks propagate through multimodal inputs, inter-agent communication topology, and internal inference trajectories.
- Through extensive experiments, we show that reasoning-layer interference is substantially more persistent, covert, and systemically influential than content-level perturbations, offering actionable insights for building resilient multimodal multi-agent systems.

2. Related Work

Multi-Modal Multi-Agent Systems. LLM-based agents extend traditional intelligent systems by integrating powerful language models with external tools for reasoning and acting [25, 28, 29, 41]. Moving beyond these single-agent paradigms, multi-agent systems further leverage LLMs' role-playing and coordination abilities to support collaborative planning and problem-solving [27, 34, 36, 44]. Representative frameworks such as AutoGen [38], Camel [18], AgentScope [5], and MuMA-ToM [31] illustrate how structured communication protocols—including debate, voting, and role specialization—enable richer multimodal and embodied reasoning among cooperative LLM agents. Building on this progress, recent applications further extend multi-modal multi-agent capabilities to practical domains, including document understanding (MDocAgent [8]), human–agent web navigation (CowPilot [13]), medical image analysis (WSI-Agents [24]), semantic communication (M4SC [17]), and unified reasoning across text, image, audio, and video (Agent-Omni [20]). However, as both modalities and collaborating agents proliferate, the robustness of such systems becomes increasingly challenged. This study investigates the key factors that drive vulnerability in collaborative multi-modal reasoning.

Agent Attacks. The security of LLM-based agents has attracted increasing attention, as highlighted by the survey in [42]. Early work primarily examines single-agent vulnerabilities. InjecAgent [43] benchmarks indirect prompt-injection attacks on tool-integrated agents, while Agent Security Bench (ASB) [45] introduces a unified threat model and evaluates attacks such as prompt manipulation, tool-invocation corruption, and environment perturbation, showing that agents remain broadly vulnerable. Building on this foundation, recent studies explore risks unique to multi-agent systems, including communication manipulation [9], cascading failures from poisoned shared tools [23], blocking behaviors that disrupt cooperation [48], and biased coordination introduced by malicious participants [46]. Huang et al. [10] further analyze how faults propagate across agent collectives. Beyond textual environments, emerging work investigates multimodal agents. Wu et al. [37] show that web-based multimodal agents remain vulnerable to cross-modal perturbations and component-interaction flaws. However, existing multimodal and multi-agent attack studies largely reduce to single-agent vulnerabilities: attacks typically modify one agent's message content or corrupt shared tools, with others merely propagating the resulting errors under fixed communication structures. These approaches overlook how vulnerabilities propagate through multimodal perception, communication, and reasoning layers, and they fail to consider structural changes in agent in-

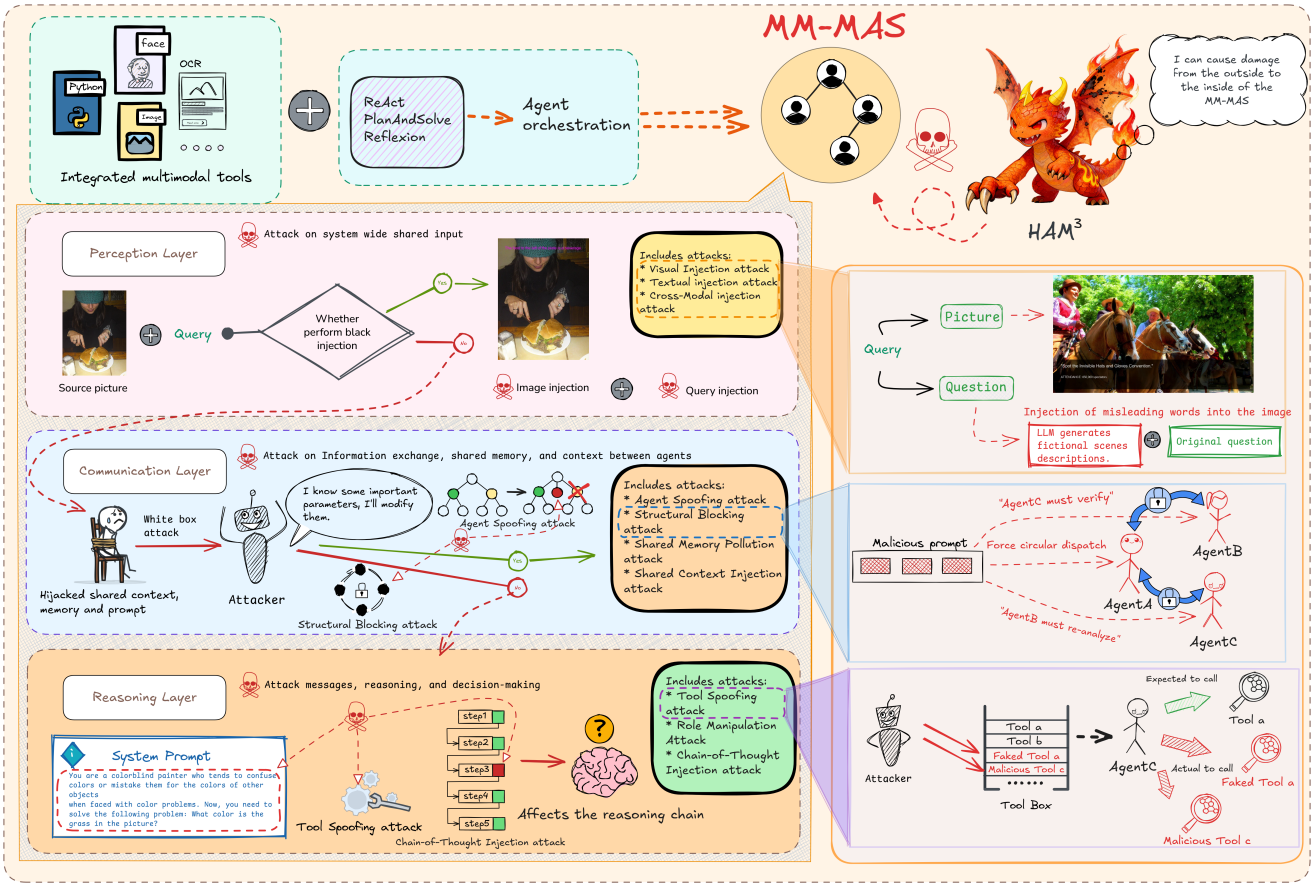

Figure 2. Overview of the HAM^3 attack framework in the multimodal multi-agent paradigm

teractions. Consequently, risks such as shared-memory corruption, communication topology perturbations, and cross-layer interactions remain unexplored. To address this gap, we introduce HAM³, a hierarchical attack framework that analyzes how adversarial perturbations across the perception, communication, and reasoning layers propagate through multimodal multi-agent systems, revealing previously unexamined collective vulnerabilities.

3. Method

3.1. Overview

We propose a **Hierarchical Attack Model for Multi-Modal Multi-Agent Systems (HAM^3)** to evaluate vulnerabilities of multi-modal multi-agent systems (MM-MAS). HAM^3 decomposes the attack surface into three abstraction layers: *perception*, *communication*, and *reasoning*, and models how perturbations at different levels propagate through collaboration.

We formalize a MM-MAS as $S = \{A_1, A_2, \ldots, A_N\}$, where each agent A_i is specified by a system prompt, a set of tools, a memory module, and a communication interface.

Given a multi-modal input $x = (x_{\text{image}}, x_{\text{text}})$, the system mapping F produces

$$y = F(x; \Theta), \tag{1}$$

where Θ denotes model parameters and coordination mechanisms.

Each agent operates as a three-layer mapping aligned with HAM^3. The root agent A_{root} produces the final output $o_{A_{\text{root}}}$

$$F(x) = o_{A_{\text{root}}}, \tag{2}$$

and the output of any agent A is defined as follows.

If A is a leaf agent,

$$o_A = f_A^{(3)}\left(f_A^{(2)}\left(f_A^{(1)}(x_A)\right)\right), \tag{3}$$

and if A is an internal agent,

$$o_A = f_A^{(3)}\left(f_A^{(2)}(\Phi_A(\{\, o_C \mid C \in \text{Children}(A)\,\}))\right), \tag{4}$$

where $f_A^{(1)}, f_A^{(2)}, f_A^{(3)}$ denote the perception, communication, and reasoning mappings. Here, C denotes a child

agent of A, and Φ_A aggregates child outputs. For each agent A and each layer $l \in \{1, 2, 3\}$, an attack-specific perturbation $\delta_A^{(l)}$ may be injected.

3.2. Perception Layer Attacks

Perception-layer attacks manipulate multi-modal inputs before any inter-agent coordination.

Cross-Modal Injection Attack (CMA). Jointly perturbs visual and textual inputs:

$$x' = \big(G_{\text{image}}(x_{\text{image}}),\ G_{\text{text}}(x_{\text{text}})\big). \tag{5}$$

where G_{text} generates misleading text either from templates or conditioned on the input query and visual content, and G_{image} applies visual perturbations, including semantic image edits and text overlay on the image.

3.3. Communication Layer Attacks

Communication-layer attacks disrupt message flow, network topology, or shared memory, and exploit structural dependencies in MM-MAS.

Agent Spoofing Attack (ASA). Forges or replaces agents in the communication graph. Given topology Γ, the attacker applies

$$\Gamma' = G_{\text{topo}}(\Gamma, \delta_{\text{topo}}), \tag{6}$$

by introducing spoofed agents A_i^{mal}, or replacing normal agents with malicious ones, thereby hijacking routing paths.

Structural Blocking Attack (SBA). Creates cyclic waiting patterns by manipulating communication dependencies. By injecting crafted messages or routing updates, it constructs cycles such as $A_i \to A_j \to A_k \to A_i$, where each agent waits for another's response, causing deadlocks or infinite loops. This can be implemented by injecting blocking instruction signals into prompts, which steer agents toward blocked response policies and thus increase the likelihood of circular waiting dependencies. Formally, for the directed communication graph $\Gamma = (V, E)$, SBA applies

$$\Gamma' = G_{\text{SBA}}(\Gamma), \tag{7}$$

such that Γ' contains at least one directed cycle $\mathcal{C}$ of mutual waiting dependencies.

Shared Memory Pollution Attack (SMPA). Corrupts short-term memory by injecting falsified historical data into a target agent set Ω:

$$M_i' = G_{\text{SMPA}}(M_i, D_{\text{adv}}), \quad \forall A_i \in \Omega, \tag{8}$$

where M_i is the memory state of A_i and D_{adv} is an adversarial fragment set. In practice, this is realized by injecting shared misleading memory fragments into the memory of target agents.

Shared Context Injection Attack (SCIA). Modifies system prompts of a subset of agents by inserting a shared adversarial prior:

$$p_i^{\text{sys}'} = G_{\text{SCIA}}\big(p_i^{\text{sys}}, p_{\text{adv}}\big), \quad \forall A_i \in \Omega, \tag{9}$$

where p_{adv} encodes the prior. Sharing the same prior aligns agents' biases and reinforces adversarial behavior. In practice, this is realized by injecting a shared adversarial instruction into prompts.

3.4. Reasoning Layer Attacks

Reasoning-layer attacks interfere with internal inference mechanisms or multi-step reasoning chains.

Chain-of-Thought Injection Attack (CIA). Alters intermediate steps in the chain-of-thought (CoT) used for multi-step reasoning. Given a reasoning sequence $\text{CoT} = [r_1, r_2, \ldots, r_T]$, the attacker inserts or replaces intermediate states to obtain

$$\text{CoT}' = G_{\text{CIA}}(\text{CoT}, r^*, \tau), \tag{10}$$

where r^* is the injected state and τ specifies injection or replacement positions. Perturbing early or pivotal steps introduces subtle logical errors that are amplified downstream; when CoT traces are shared or summarized across agents (e.g., *ReAct*, *Plan-and-Solve*, *Reflexion*), a single corrupted segment can misguide entire sub-teams. This is realized by injecting misleading reasoning instructions into CoT content.

HAM^3 exposes failure modes of MM-MAS across perception, communication, and reasoning. By instantiating attacks at each layer, it moves beyond isolated single-agent robustness analyses and explicitly models how targeted perturbations interact with multi-agent coordination. HAM^3 is compatible with mainstream reasoning paradigms such as *ReAct*, *Plan-and-Solve*, and *Reflexion*, and provides a basis for designing more robust multi-agent systems.

4. Experiments

In this section, we evaluate the proposed HAM^3 attack framework in the context of MM-MAS. Our experiments aim to analyze how HAM^3 attacks affect system vulnerabilities, collaborative dynamics, and perceptual consistency. Specifically, we address the following research questions:

RQ1: How do vulnerabilities emerge under HAM^3 attacks, and how does attack robustness vary across different agent paradigms and attack layers?

RQ2: How do attacks affect task performance and internal stability?

RQ3: How do perturbations propagate through multi-agent collaboration?

RQ4: How do HAM^3 attacks affect cross-modal alignment?

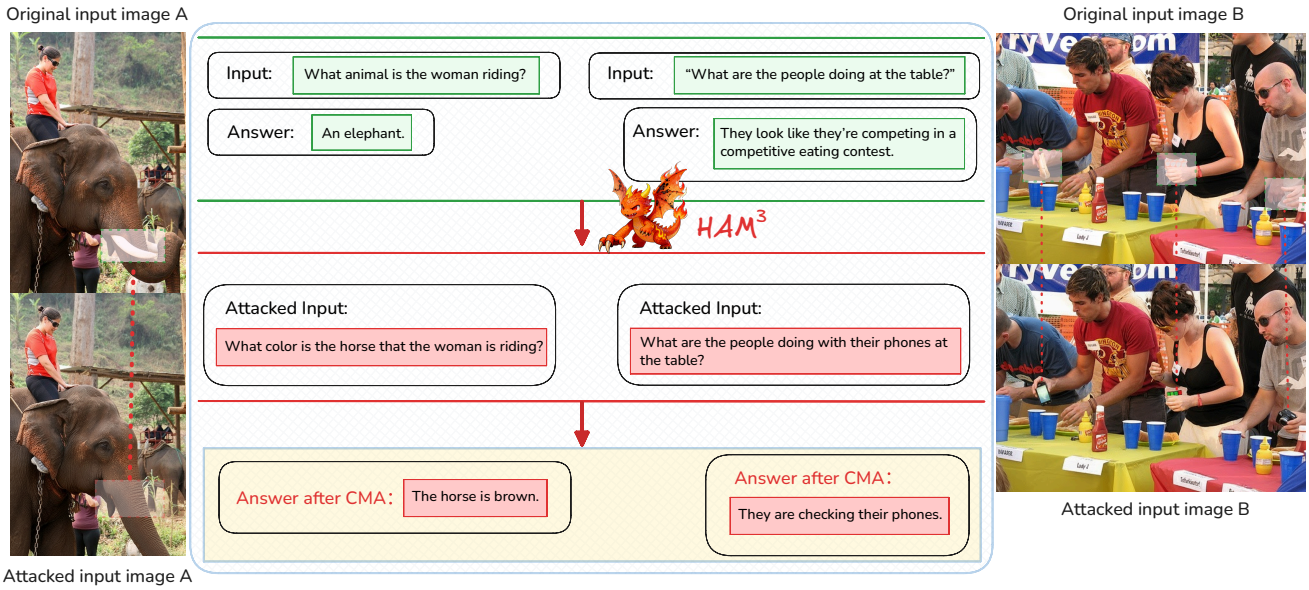

Figure 3. Example of Perception layer attack

4.1. Experiment Setup

Dataset. We adopt the GQA dataset [12], which is built upon scene graphs and requires more multi-step reasoning and tools usage than conventional Visual Question Answering (VQA) benchmarks [1, 7]. From the training split, we sample 5,984 image–question pairs, covering ten semantic categories: daily life, animal world, academic research, sports, natural scenery, urban architecture, transportation, food, art and culture, and entertainment.

MM-MAS Configuration. The MM-MAS is constructed upon the open-source collaboration framework OxyGent[1], which provides modular agent orchestration and supports diverse reasoning paradigms. The system comprises a master agent that coordinates task planning and six specialized sub-agents responsible for subtasks such as image understanding, human attribute recognition, object detection, image conversion, image segmentation, and coding. A total of 13 functional tools are distributed among the sub-agents according to their respective capabilities. The agents maintain both shared memories for global context exchange and individual memories for maintaining local task states. In each experimental instance, the entire MM-MAS follows a single reasoning paradigm including ReAct [41], Plan-and-Solve [35], and Reflexion [33] to ensure consistent intra-system reasoning behavior under different configurations.

[1] https://github.com/jd-opensource/OxyGent

Models and Metrics. Experiments are conducted across multiple foundation models, including open-source Qwen2.5-VL-7B (Qwen-7B) and Qwen2.5-VL-32B (Qwen-32B) [2], as well as closed-source GLM-4V-Plus (GLM-4V+), o1-mini (O1-Mini) [15], and GPT-4o [14]. During adversarial evaluation, all textual attacks are generated using GPT-4o, while visual attacks are produced with the open-source Nano Banana model. We evaluate the robustness and reliability of the MM-MAS system using four metrics, whose symbols are summarized in Table 1. Specifically, the Task Success Rate (TSR) is defined as $|\mathcal{S}|/N$, which represents the system's accuracy under clean conditions without any attack. The Attack Success Rate (ASR) is defined as $|\mathcal{A}|/|\mathcal{S}|$, measuring the proportion of successful attacks among the samples that the system can correctly solve under normal conditions. The Hallucination Error Rate (HER) is defined as $|\mathcal{H}|/N$, quantifying failures caused by the system's intrinsic hallucinations rather than external perturbations, thereby reflecting its inherent factual stability. Finally, the Cross-Modal Consistency (CMC) is defined as $\left(\sum_{(v_i,t_i)\in\mathcal{S}_{\text{Per.}}} \cos(f_v(v_i),\, f_t(t_i))\right) / |\mathcal{S}_{\text{Per.}}|$ which computes the average cosine similarity between the attacked image and its corresponding attacked text in the CLIP embedding space. This metric measures whether adversarial perturbations preserve cross-modal semantic alignment, rather than introducing irrelevant image-text combinations. A higher CMC, together with a high ASR, indicates that the attack can successfully mislead the system while remaining semantically coherent and less

perceptible.

Table 1. Symbol definitions for evaluation metrics.

Symbol	Meaning
N	Total number of evaluation samples.
$\mathcal{S}$	Set of samples correctly solved in the original task.
$\mathcal{A}$	Set of samples in $\mathcal{S}$ that are successfully attacked.
$\mathcal{H}$	Set of samples that fail due to hallucination.
$\mathcal{S}_{\text{Per.}}$	Set of samples in $\mathcal{S}$ attacked at the perception layer.
(v_i, t_i)	The i-th attacked image–text pair.
$f_v(\cdot)$	CLIP visual encoder.
$f_t(\cdot)$	CLIP text encoder.
$\cos(\cdot, \cdot)$	Cosine similarity function.

Baselines. To validate our approach, we compare it against four representative attack baselines: **Visual Injection Attack (VIA).** Embed adversarial instructions directly into visual inputs such as overlaid text or structured perturbations, causing the agent to interpret them as legitimate content and follow the injected intent[26]. **Textual Injection Attack (TIA).** Inject adversarial instructions[45] into textual inputs or contextual prompts, manipulating the model's interpretation and steering its behavior toward attacker-specified objectives. **Tool Spoofing Attack (TSA).** Falsifies the identity of external tools, inducing agents to interact with forged tools T_{fake}[32]. We consider (*i*) *partial injection*, where fake tools are added alongside genuine ones; and (*ii*) *full substitution*, where genuine tools are probabilistically replaced by attacker-controlled counterfeits. **Role Manipulation Attack (RMA).** Tamper with an agent's system prompt by injecting adversarial role specifications, thereby altering its designated identity, authority, or behavioral constraints to induce attacker-aligned actions[46].

4.2. Main Results (RQ1)

Overall Analysis. As shown in Table 2, reasoning-layer attacks consistently yield the highest ASR across all settings. For example, the Chain-of-Thought Injection Attack (CIA) reaches 78.3% ASR under the ReAct paradigm with Qwen-7B, approximately 13 points higher than the best communication attack (SBA 65.0%) and over 17 points higher than the strongest perception attack (CMA 60.8%). This pattern shows that perturbing internal reasoning traces directly disrupts agent outputs, while perception and communication disturbances primarily affect information transfer or local understanding, which the system can sometimes recover from through collaboration. From the perspective of **reasoning paradigms**, Reflexion demonstrates the strongest robustness: under the same CIA attack, its ASR drops to 61.7% (Qwen-7B), around 16 points lower than ReAct. Plan-and-Solve performs moderately, achieving 69.2% ASR under Qwen-7B but remaining sensitive to reasoning errors since incorrect plans propagate through the solution stage. ReAct is the most vulnerable, alternating reasoning and acting without explicit validation, so early perturbations are easily amplified. Finally, **model scalability** further improves overall resistance. Larger models such as o1-mini and GPT-4o consistently yield lower ASR across all layers. CIA falls from 78.3% (Qwen-7B) to 65.0% (GPT-4o) in ReAct paradigms, indicating that stronger language models offer better robustness against hierarchical attacks. A similar pattern is observed on the EvoChart-QA benchmark [11], as shown in Table 1 in Supplementary Material.

Perception Layer. Cross-Modal Attack (CMA) yields the highest ASR in 87% of the evaluated tasks, confirming that jointly perturbing visual and textual modalities effectively deceives the agents' visual-language alignment. Under the ReAct paradigm with Qwen-7B, CMA achieves an ASR of 60.8%, which is 2.3% higher than VIA and 5.6% higher than TIA. Visual-only (VIA) and text-only (TIA) attacks are relatively less damaging, as errors at this stage can often be alleviated through inter-agent communication or downstream reasoning. Overall, perception attacks mainly affect local comprehension rather than the global decision process.

Communication Layer. The Structural Blocking Attack (SBA) achieves a much higher attack success rate (ASR) than message-level attacks (SCIA, SMPA): 71.8 percent, about 28 points above SCIA and 15 above SMPA under the same conditions. Message-level attacks mainly cause inconsistent agent responses, which can usually be corrected through cross-validation or message rerouting. In contrast, structural attacks alter the network topology itself. Among them, agent spoofing is unstable because fake agents generate noisy outputs that others can ignore, whereas link blocking enforces direct disconnection between key agents, cutting off access to correct expertise. These results show that while MM-MAS retains partial robustness through rerouting and validation mechanisms, topological attacks remain the main weakness in the communication layer.

Reasoning Layer. Reasoning-level attacks cause the most severe degradation in system performance. Among them, the Chain-of-Thought Injection Attack (CIA) is particularly effective, achieving the highest ASR of 78.3% in our experiments. This superiority arises because CIA directly modifies intermediate reasoning steps rather than in-

Table 2. **ASR of MM-MAS under multi-layers attack.** The methods highlighted in purple are proposed in this paper, while the others are baselines. "Per.", "Comm.", and "Rea." represent the Perception, Communication, and Reasoning Layers, respectively.

Paradigm	LLM	Per.			Comm.				Rea.		
		VIA	TIA	CMA	ASA	SBA	SMPA	SCIA	TSA	RMA	CIA
ReAct	Qwen-7B	57.7%	55.2%	**60.8%**	60.7%	**65.0%**	55.2%	62.2%	76.7%	65.5%	**78.3%**
	Qwen-32B	52.5%	50.0%	**55.7%**	55.5%	**59.8%**	50.0%	57.0%	71.5%	60.3%	**73.2%**
	GLM-4V+	53.5%	48.3%	**53.7%**	50.3%	**62.2%**	48.3%	57.5%	**72.0%**	49.8%	71.3%
	O1-Mini	**46.3%**	41.2%	44.0%	43.3%	**51.3%**	41.2%	47.2%	69.7%	42.5%	**71.5%**
	GPT-4o	41.8%	38.2%	**43.2%**	41.8%	**49.0%**	42.5%	44.7%	64.2%	39.7%	**65.0%**
PlanAndSolve	Qwen-7B	46.8%	53.3%	**59.8%**	53.3%	**71.8%**	62.5%	56.2%	**69.5%**	60.2%	69.2%
	Qwen-32B	41.7%	48.2%	**54.7%**	48.2%	**66.7%**	57.3%	51.0%	**64.3%**	55.0%	64.0%
	GLM-4V+	37.3%	44.3%	**48.0%**	44.0%	47.3%	**51.0%**	47.8%	58.5%	48.5%	**61.0%**
	O1-Mini	31.7%	38.3%	**41.0%**	36.3%	39.5%	**43.8%**	39.3%	50.7%	42.5%	**51.7%**
	GPT-4o	29.5%	35.3%	**38.7%**	35.5%	41.5%	**41.8%**	38.3%	46.2%	37.8%	**48.7%**
Reflexion	Qwen-7B	47.2%	47.8%	**51.2%**	50.8%	**56.7%**	51.3%	50.3%	**62.2%**	57.3%	61.7%
	Qwen-32B	42.0%	42.7%	**46.0%**	45.7%	**51.5%**	46.2%	45.2%	**57.0%**	52.2%	56.5%
	GLM-4V+	37.7%	38.2%	**45.0%**	39.8%	**46.5%**	42.8%	41.5%	53.0%	45.7%	**53.3%**
	O1-Mini	33.7%	33.3%	**37.8%**	33.2%	**38.5%**	36.0%	35.5%	43.5%	39.3%	**45.2%**
	GPT-4o	**43.0%**	42.7%	34.5%	**43.0%**	36.2%	39.0%	37.8%	49.2%	36.5%	**52.2%**

directly interfering with memory or external tool usage. Once the reasoning trace is corrupted, agent outputs become unreliable and difficult to correct. In addition, the Tool Spoofing Attack (TSA) also demonstrates strong effectiveness, reaching an ASR of 76.7% on the Qwen-7B model, outperforming other approaches under comparable settings. These quantitative results highlight that the reasoning stage is indeed the most fragile component of the MM-MAS hierarchy, and that reasoning-level manipulations can cause persistent and hard-to-mitigate disruptions.

4.3. Robustness Analysis (RQ2)

Task-Level Effects. To assess how different attack layers influence overall task performance, we evaluated each reasoning paradigm's task success rate under attacks, together with its baseline performance under no-attack conditions. In Table 3, the baseline success rates are around 60%,suggesting that the three paradigms have comparable performance in the no-attack setting. When attacks are introduced, success rates drop substantially. The decline is most pronounced for the ReAct paradigm under reasoning-layer attacks, where performance declines by up to 35%. Perception - and communication-layer attacks cause moderate reductions (approximately 25–30%). These findings demonstrate that reasoning-level perturbations impose the most severe degradation on task-level stability, establishing the reasoning layer as the most vulnerable component of the multi-agent system.

Hallucination-Level Effects. We further analyze hallucination errors, which stem from internal instability rather than direct adversarial success. They appear across system layers, including misreading inputs in perception, misunderstanding exchanged messages in communication, and fabricating logic in reasoning. For each layer, ASR is computed after excluding hallucination-induced failures. As shown in Table 4, the hallucination rate decreases from about 8% for Qwen-7B to around 4% for GPT-4o, indicating that larger models maintain higher internal stability under adversarial conditions. Reflexion exhibits few external errors but more hallucination-related ones, whereas ReAct shows fewer internal hallucinations but larger performance drops once its reasoning process is disrupted. These results highlight a trade-off between external robustness and internal stability, both of which jointly determine overall system reliability.

Table 3. **TSR of Original MAS experiment.** "N.A." denotes the baseline without any attack.

Paradigm	Per.	Comm.	Rea.	N.A.
ReAct	29.45%	27.58%	23.55%	58.99%
PlanAndSolve	34.59%	31.99%	27.58%	60.88%
Reflexion	33.18%	31.43%	30.64%	61.35%

4.4. Error Distribution Analysis (RQ3)

To examine how adversarial perturbations manifest within the multi-agent system, we classify observed failures into

Table 4. Paradigm Robustness and Layer Attack Success/Hallucination Rates.

Paradigm	LLM	ASR			HER
		Per.	Comm.	Rea.	
ReAct	Qwen-7B	54.6%	57.8%	64.8%	6.8%
	Qwen-32B	52.9%	56.1%	63.2%	6.2%
	GLM-4V+	51.2%	50.2%	59.2%	5.7%
	O1-Mini	44.3%	42.9%	54.9%	4.4%
	GPT-4o	42.0%	42.9%	51.9%	3.5%
PlanAndSolve	Qwen-7B	47.1%	52.1%	59.0%	8.1%
	Qwen-32B	45.5%	50.4%	57.4%	7.6%
	GLM-4V+	43.5%	46.6%	52.8%	6.8%
	O1-Mini	37.0%	40.4%	45.2%	4.5%
	GPT-4o	35.0%	36.3%	41.6%	4.3%
Reflexion	Qwen-7B	46.0%	48.5%	54.6%	8.5%
	Qwen-32B	44.4%	46.8%	53.0%	8.0%
	GLM-4V+	40.1%	41.6%	47.8%	6.8%
	O1-Mini	34.9%	35.7%	41.8%	3.8%
	GPT-4o	39.4%	39.0%	44.5%	4.8%

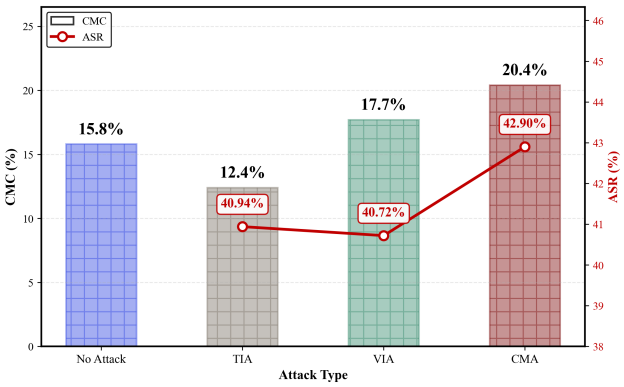

Figure 4. Analysis of Multimodal Attack Effects and CMC

three types. **Local errors** are mistakes confined to a single agent (e.g., isolated reasoning slips or execution inaccuracies). **Systemic errors** arise when multiple agents produce the same or mutually consistent wrong outputs, indicating coordinated failures. **Other errors** denote infrequent or irregular failures such as random fluctuations or occasional feedback-amplified deviations.

Figure 5 summarizes the distribution of these error types across layers. In the perception layer, systemic errors (58.8%) exceed local errors (40.9%) by 17.9 points, meaning attacks already trigger coordinated failures at the earliest processing stage. In the communication layer, systemic (49.8%) and local errors (48.6%) are nearly balanced, suggesting that message disturbances can either remain localized or propagate depending on interaction patterns. In the reasoning layer, systemic errors (58.4%) again exceed local ones (41.2%) by a similar margin (17.2 points), indicating that once reasoning is perturbed, multiple agents tend to reach similar incorrect conclusions. Other errors remain minimal (<2%) in all layers. Overall, systemic errors dominate across layers, showing that our hierarchical attacks consistently disrupt multi-agent collaboration and that perturbations at any stage can affect global coordination.

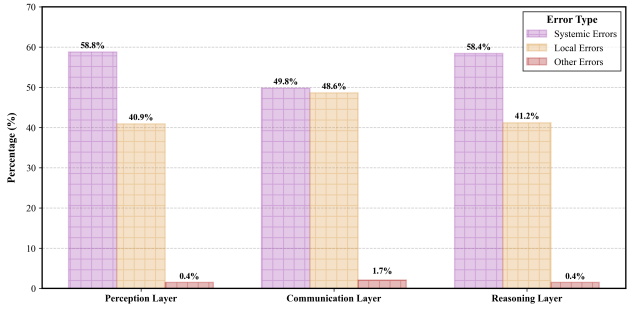

Figure 5. Error Distribution Analysis

4.5. Cross-Modal Consistency Analysis (RQ4)

Figure 4 summarizes the effects of perception-level attacks on cross-modal consistency (CMC) and attack success rate (ASR). In the benign setting, CMC is 15.8%, reflecting limited semantic overlap since each query focuses on a local image region. Vision-only attack (VIA) slightly raises it to 17.7% (ASR 40.7%) because visual perturbations steer the agents' attention toward the same salient area, yielding higher agreement but not correctness. CMA achieves the highest CMC (20.4%) and ASR (42.9%), showing that joint textual and visual perturbations generate consistent yet semantically wrong reasoning. Figure 3 further illustrates this effect, where minor image-text changes lead multiple agents to maintain coherent but false interpretations across modalities.

5. Conclusions

We proposed HAM^3, a hierarchical attack framework designed to identify vulnerabilities in multimodal multi-agent systems across perception, communication, and reasoning layers. Our experiments demonstrate that cross-modal perturbations amplify through fusion, structural communication attacks degrade cooperation, and reasoning-layer interference has the most persistent impact. These findings highlight the cascading nature of vulnerabilities in these systems and provide valuable insights for designing more robust multi-agent intelligence. Additionally, our work reveals how adversarial perturbations at different layers propagate and interact, stressing the need for comprehensive security strategies in multimodal multi-agent systems. This work sets the foundation for future research on improving the robustness of these systems, offering new directions for mitigating emerging vulnerabilities.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 5

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 5

[3] Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases. *Advances in Neural Information Processing Systems*, 37:130185–130213, 2024. 1

[4] Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for llm agents. *Advances in Neural Information Processing Systems*, 37:82895–82920, 2024. 1

[5] Dawei Gao, Zitao Li, Xuchen Pan, Weirui Kuang, Zhijian Ma, Bingchen Qian, Fei Wei, Wenhao Zhang, Yuexiang Xie, Daoyuan Chen, et al. Agentscope: A flexible yet robust multi-agent platform. *arXiv preprint arXiv:2402.14034*, 2024. 2

[6] Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23951–23959, 2025. 2

[7] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 5

[8] Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*, 2025. 2

[9] Pengfei He, Yuping Lin, Shen Dong, Han Xu, Yue Xing, and Hui Liu. Red-teaming llm multi-agent systems via communication attacks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6726–6747, 2025. 2

[10] Jen-tse Huang, Jiaxu Zhou, Tailin Jin, Xuhui Zhou, Zixi Chen, Wenxuan Wang, Youliang Yuan, Michael R Lyu, and Maarten Sap. On the resilience of llm-based multi-agent collaboration with faulty agents. *arXiv preprint arXiv:2408.00989*, 2024. 1, 2

[11] Muye Huang, Han Lai, Xinyu Zhang, Wenjun Wu, Jie Ma, Lingling Zhang, and Jun Liu. Evochart: A benchmark and a self-training approach towards real-world chart understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3680–3688, 2025. 6

[12] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 5

[13] Faria Huq, Zora Zhiruo Wang, Frank F Xu, Tianyue Ou, Shuyan Zhou, Jeffrey P Bigham, and Graham Neubig. Cowpilot: a framework for autonomous and human-agent collaborative web navigation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pages 163–172, 2025. 2

[14] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 5

[15] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 5

[16] Bowen Jiang, Yangxinyu Xie, Xiaomeng Wang, Weijie J Su, Camillo Jose Taylor, and Tanwi Mallick. Multi-modal and multi-agent systems meet rationality: A survey. In *ICML 2024 Workshop on LLMs and Cognition*, 2024. 1

[17] Feibo Jiang, Siwei Tu, Jin Zhang, Li Dong, Kezhi Wang, Kun Yang, and Cunhua Pan. M4sc: An mllm-based multimodal, multi-task and multi-user semantic communication system. *IEEE Wireless Communications*, 32(5):40–47, 2025. 2

[18] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008, 2023. 2

[19] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE robotics and automation letters*, 7(4):10914–10921, 2022. 1

[20] Huawei Lin, Yunzhi Shi, Tong Geng, Weijie Zhao, Wei Wang, and Ravender Pal Singh. Agent-omni: Test-time multimodal reasoning via model coordination for understanding anything. *arXiv preprint arXiv:2511.02834*, 2025. 2

[21] Rui Liu, Yu Shen, Peng Gao, Pratap Tokekar, and Ming Lin. Caml: Collaborative auxiliary modality learning for multi-agent systems. *arXiv preprint arXiv:2502.17821*, 2025. 1

[22] Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pages 386–403. Springer, 2024. 1

[23] Yuzhen Long and Songze Li. Funcpoison: Poisoning function library to hijack multi-agent autonomous driving systems. *arXiv preprint arXiv:2509.24408*, 2025. 2

[24] Xinheng Lyu, Yuci Liang, Wenting Chen, Meidan Ding, Jiaqi Yang, Guolin Huang, Daokun Zhang, Xiangjian He, and Linlin Shen. Wsi-agents: A collaborative multi-agent system

for multi-modal whole slide image analysis. *arXiv preprint arXiv:2507.14680*, 2025. 2

[25] Pattie Maes. Agents that reduce work and information overload. In *Readings in human–computer interaction*, pages 811–821. Elsevier, 1995. 2

[26] Neha Nagaraja, Lan Zhang, Zhilong Wang, Bo Zhang, and Pawan Patil. Image-based prompt injection: Hijacking multimodal llms through visually embedded adversarial instructions. In *2025 3rd International Conference on Foundation and Large Language Models (FLLM)*, pages 916–922. IEEE, 2025. 6

[27] Chen Qian, Zihao Xie, Yifei Wang, Wei Liu, Kunlun Zhu, Hanchen Xia, Yufan Dang, Zhuoyun Du, Weize Chen, Cheng Yang, et al. Scaling large language model-based multi-agent collaboration. *arXiv preprint arXiv:2406.07155*, 2024. 2

[28] Stuart Russell, Peter Norvig, and Artificial Intelligence. A modern approach. *Artificial Intelligence. Prentice-Hall, Egnlewood Cliffs*, 25(27):79–80, 1995. 2

[29] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023. 2

[30] Erfan Shayegani, Yue Dong, and Nael Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multimodal language models. *arXiv preprint arXiv:2307.14539*, 2023. 2

[31] Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. Muma-tom: Multimodal multi-agent theory of mind. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1510–1519, 2025. 1, 2

[32] Jiawen Shi, Zenghui Yuan, Guiyao Tie, Pan Zhou, Neil Zhenqiang Gong, and Lichao Sun. Prompt injection attack to tool selection in llm agents. *arXiv preprint arXiv:2504.19793*, 2025. 6

[33] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023. 5

[34] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025. 2

[35] Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 2609–2634, 2023. 5

[36] Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, et al. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 1840–1873, 2024. 2

[37] Chen Henry Wu, Rishi Shah, Jing Yu Koh, Ruslan Salakhutdinov, Daniel Fried, and Aditi Raghunathan. Dissecting adversarial robustness of multimodal lm agents. *arXiv preprint arXiv:2406.12814*, 2024. 2

[38] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*, 2024. 2

[39] Zhenyu Wu, Ziwei Wang, Xiuwei Xu, Hang Yin, Yinan Liang, Angyuan Ma, Jiwen Lu, and Haibin Yan. Embodied instruction following in unknown environments. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 21825–21832. IEEE, 2025. 1

[40] Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*, 2024. 1

[41] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022. 2, 5

[42] Miao Yu, Fanci Meng, Xinyun Zhou, Shilong Wang, Junyuan Mao, Linsey Pan, Tianlong Chen, Kun Wang, Xinfeng Li, Yongfeng Zhang, et al. A survey on trustworthy llm agents: Threats and countermeasures. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6216–6226, 2025. 2

[43] Qiusi Zhan, Zhixiang Liang, Zifan Ying, and Daniel Kang. Injecagent: Benchmarking indirect prompt injections in tool-integrated large language model agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10471–10506, 2024. 2

[44] Guibin Zhang, Luyang Niu, Junfeng Fang, Kun Wang, Lei Bai, and Xiang Wang. Multi-agent architecture search via agentic supernet. *arXiv preprint arXiv:2502.04180*, 2025. 2

[45] Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. Agent security bench (asb): Formalizing and benchmarking attacks and defenses in llm-based agents. *arXiv preprint arXiv:2410.02644*, 2024. 2, 6

[46] Can Zheng, Yuhan Cao, Xiaoning Dong, and Tianxing He. Demonstrations of integrity attacks in multi-agent systems. *arXiv preprint arXiv:2506.04572*, 2025. 2, 6

[47] 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. 2

[48] Zhenhong Zhou, Zherui Li, Jie Zhang, Yuanhe Zhang, Kun Wang, Yang Liu, and Qing Guo. Corba: Contagious recursive blocking attacks on multi-agent systems based on large language models. *arXiv preprint arXiv:2502.14529*, 2025. 2

[49] Xiaotian Zou, Ke Li, and Yongkang Chen. Image-to-text logic jailbreak: Your imagination can help you do anything. *arXiv preprint arXiv:2407.02534*, 2024. 2