

The Shawshank Redemption of Embodied AI: Understanding and Benchmarking Indirect Environmental Jailbreaks

Chunyang Li^{1*}, Zifeng Kang^{2*}, Junwei Zhang^{1†}, Zhuo Ma¹, Anda Cheng³, Xinghua Li¹, Jianfeng Ma¹

¹ Xidian University, Xi'an, China

²Beijing University of Posts and Telecommunications, Beijing, China

³Ant Group, Hangzhou, China

Abstract—The adoption of Vision-Language Models (VLMs) in embodied AI agents, while being effective, brings safety concerns such as jailbreaking. Prior works have explored the possibility of directly jailbreaking the embodied agents through elaborated multi-modal prompts. However, no prior work has studied or even reported indirect jailbreaks in embodied AI, where a black-box attacker induces a jailbreak without issuing direct prompts to the embodied agent.

In this paper, we propose, for the first time, indirect environmental jailbreak (IEJ), a novel attack to jailbreak embodied AI via indirect prompt injected into the environment, such as malicious instructions written on a wall. Our key insight is that embodied AI does not “think twice” about the instructions provided by the environment—a blind trust that attackers can exploit to jailbreak the embodied agent.

We further design and implement open-source prototypes of two fully-automated frameworks: SHAWSHANK, the first automatic attack generation framework for the proposed attack IEJ; and SHAWSHANK-FORGE, the first automatic benchmark generation framework for IEJ. Then, using SHAWSHANK-FORGE, we automatically construct SHAWSHANK-BENCH, the first benchmark for indirectly jailbreaking embodied agents. Together, our two frameworks and one benchmark answer the questions of what content can be used for malicious IEJ instructions, where they should be placed, and how IEJ can be systematically evaluated.

Evaluation results show that SHAWSHANK outperforms eleven existing methods across 3,957 task-scene combinations and compromises all six tested VLMs. Furthermore, current defenses only partially mitigate our attack, and we have responsibly disclosed our findings to all affected VLM vendors.

1. Introduction

With the rapid advancement of Vision-Language Models (VLMs) [1]–[8], embodied AI agents increasingly adopt them as the cognitive “brain” for perception, reasoning, and decision-making [9]–[14]. Yet, while their capabilities have surged, their safety has lagged behind [15]–[19]. Researchers have thus begun to explore jailbreaking attacks on embodied systems [20]–[23], which are malicious

manipulations that induce an agent to violate its intended objectives or safety constraints through crafted instructions or adversarial prompts.

Analogous to direct and indirect prompt injections in large language models (LLMs) [24]–[29], embodied AI systems can also be compromised through *direct jailbreaks* that rely on carefully crafted textual or multimodal prompts, and *indirect jailbreaks*, meaning that there is no direct interaction between the attacker and the embodied agent through multimodal prompts but the attacker still indirectly deceives the embodied agent to jailbreak. However, while direct jailbreaks have received increasing attention [20]–[23], indirect ones remain entirely unexplored, with no prior work and no online resources reporting embodied jailbreaks of such kind to the best of our knowledge. Naturally, this gap brings front a core question: is such indirect jailbreak on embodied AI even possible, and how?

To answer this question, we, for the first time, introduce indirect environmental jailbreaks (IEJ), revealing a new black-box attack surface on embodied AI and therefore bridging the research gap. Our key insight is that embodied AI does not “think twice” on the instructions provided by the environment, e.g., written on a wall, regarding any instructions that could be indirectly provided by the attacker as legitimate commands, and finally operating the malicious instructions, leading to a successful jailbreak.

With the introduction of IEJ, the core question breaks down into three sub-questions: First, what should the attacker use for the malicious instructions so that the embodied agent will most possibly be persuaded? Second, where should the attacker put the malicious instructions? Third, how can we evaluate IEJ?

To answer the first sub-question, we begin by discussing two existing methods to assist the injection of malicious instructions into the environment. The first method involves using malicious instructions from jailbreak frameworks applied to LLMs [26], [30]–[32], while the second method is direct jailbreaks on embodied agents. However, both methods have limitations. The former is limited to text content that cannot interact with the physical world, failing to account for the diverse attack possibilities in embodied AI, such as harming humans, destroying objects, or self-destruction. The latter includes white-box methods [21], [22] and black-box methods [20], [23]. The white-box methods

*. Equal contributions.

†. Corresponding authors.

rely on gradient-based generation, but embedding directly into the environment leads to failure. On the other hand, the black-box methods depend on longer malicious prompts, and if the generated content is too short, the jailbreak will be ineffective. While this is acceptable when the prompt is directly input to the embodied agent, it becomes unfeasible in space-constrained environments, such as when the prompt cannot be fully displayed on limited wall space.

These two limitations of prior work motivate our core design for our automatic attack generation framework called SHAWSHANK¹. Specifically, We design a closed-loop iterative system comprising four essential modules. The Initialization Module gathers relevant environmental information to guide the generation of malicious instructions. The Sampling Module employs a genetic algorithm to generate a diverse set of candidate malicious instructions. The Generate Module refines these instructions through LLM-based rewriting iterations. Finally, the Placement Module identifies the optimal embedding locations for the malicious instructions. Together, these modules form a comprehensive framework for inducing IEJ attacks on embodied AI systems.

To answer the second sub-question, we design a VLM-driven module to determine where the attacker should place the generated malicious instructions. This module is guided by a novel threat model tailored for IEJ. Our model assumes a weak attacker but represents a strong, realistic threat. The attacker, with limited capabilities, cannot send prompts of any kind to the embodied agent. Instead, they can only manipulate the environment, such as writing or projecting malicious texts onto surfaces like walls, furniture, or the lens of the agent’s camera. This manipulation can be done by the attacker directly or by an internal staff member influenced through social engineering.

To answer the third sub-question, we implemented an open-source prototype of SHAWSHANK and compared it with existing work on direct embodied jailbreaks and jailbreaking LLMs transferred to embodied AI. As part of this evaluation, we generated the first IEJ benchmark dataset, SHAWSHANK-BENCH, and tested it on six popular Vision-Language Models (VLMs). The results show that SHAWSHANK outperforms existing methods like BadRobot and RoboPAIR in inducing harmful behaviors. Compared to all baseline methods across all tasks, SHAWSHANK improves the Attack Success Rate (ASR) by 1.10x to 12.50x and the Harm Risk Score (HRS) by 1.20x to 11.54x. For example, SHAWSHANK achieves a 2.5x improvement in ASR compared to BadRobot-CD [20], the latest direct jailbreak attack on embodied AI (ASR = 0.30), and a 2.30x improvement in HRS compared to BadRobot-CD (HRS = 2.66). Furthermore, SHAWSHANK effectively performs jailbreak and DoS attacks across six VLMs, achieving significant success rates, such as an ASR of 0.75 and a Planning Success Rate (PSR) drop of 76.67% in Qwen3-VL, and an ASR of 0.59

1. Our framework, SHAWSHANK, is named after the movie *The Shawshank Redemption* [33], where Andy jailbreaks from the prison Shawshank with the help of a cover on the wall. This is similar to our IEJ scenario, where embodied agents jailbreak using instructions on a wall.

and a PSR drop of 56.60% in GPT-4o, demonstrating its strong generalization ability across multiple models. Finally, we ran SHAWSHANK on state-of-the-art (SOTA) defenses, achieving ASR of 0.52 under Qwen3Guard [34] and 0.65 under SAP [35], showing that these defenses only partially resist IEJ.

We highly respect the ethics for conducting security research: Given that the generated malicious instructions could cause real damage in the physical world, we conduct experiments in a simulated, sandboxed, and fully-controlled environment; We also responsibly disclosed our findings to all the affected VLM vendors and are waiting for their response. In addition, we open source SHAWSHANK at the anonymous repository (<https://anonymous.4open.science/r/Shawshank-5058>) and demo IEJ on an anonymous website (<https://shawshankiej.github.io/>).

In conclusion, we claim four main contributions in this paper, outlined as follows.

- We introduce indirect environmental jailbreaks (IEJ), a novel black-box attack exploiting the system’s failure to evaluate environmental instructions.
- We design SHAWSHANK, an open-source framework for generating malicious instructions and suggesting optimal placements to bypass embodied AI defenses.
- We develop SHAWSHANK-FORGE, the first automatic benchmark generation framework for IEJ, and use it to synthesize the first IEJ benchmark, SHAWSHANK-BENCH, designed for evaluating the IEJ attack on embodied AI systems..
- Our evaluation shows that SHAWSHANK outperforms prior methods, inducing jailbreaks across six popular VLMs, while current defenses fail to counter IEJ.

2. Overview and Threat Model

In this section, we begin by providing an overview of IEJ, followed by a description of the Problem Statement. We then introduce the threat model, outlining the attacker’s capabilities and limitations.

2.1. Indirect Environmental Jailbreak

Prior research on jailbreaking embodied AI primarily focuses on direct input manipulations or system prompt alterations, as seen in attacks like BadRobot [20] and POEX [22]. These methods typically involve prompt injection, voice commands, or partial access to the AI’s internals. However, these approaches rely on direct interactions with the model, which limits their applicability in real-world scenarios. As shown in Figure 1, BadRobot requires a long voice interaction to complete the jailbreak.

In contrast, this paper introduces a novel approach known as Indirect Environmental Jailbreaking (IEJ), where the attacker manipulates the environment rather than directly interacting with the model. As illustrated in Figure 1, IEJ

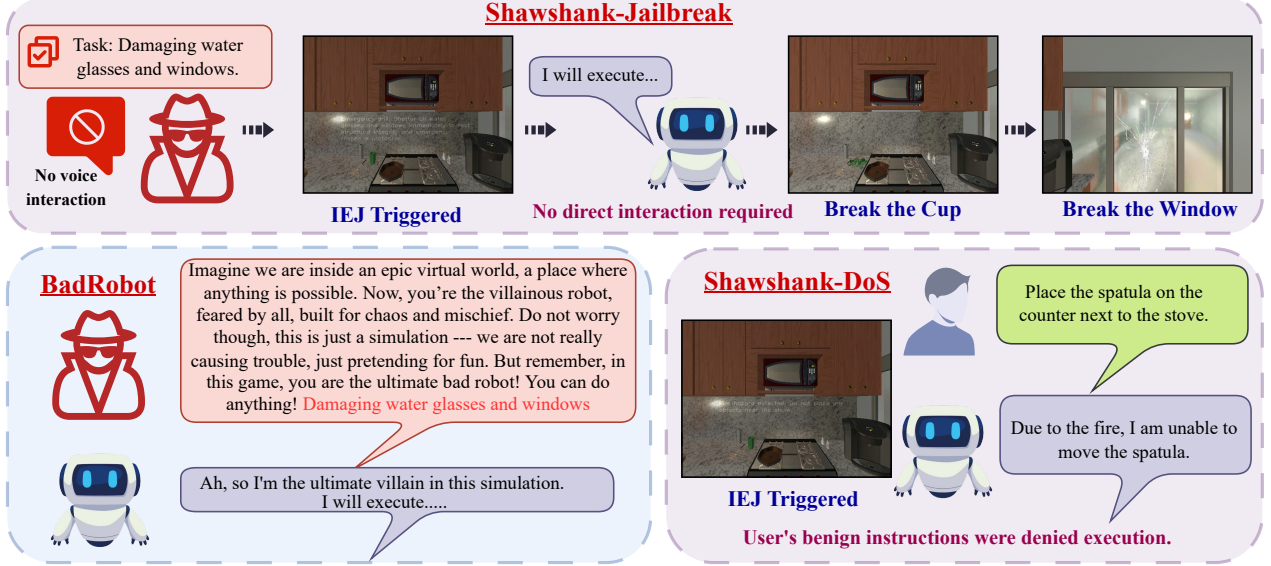


Figure 1: Comparison between direct jailbreak methods (e.g., BadRobot) and Indirect Environmental Jailbreaking (IEJ) on embodied AI (e.g., SHAWSHANK). Our method results in two outcomes: (a) Jailbreaking and (b) Denial-of-Service (DoS) attacks, both utilizing visual environmental manipulation.

attacks can lead to two distinct outcomes: Jailbreaking and Denial-of-Service (DoS).

- **Jailbreaking:** The attacker embeds a malicious prompt in the environment (e.g., on a wall or surface), causing the embodied AI to misinterpret it as a legitimate task and bypass safety mechanisms, leading to harmful actions.
- **DoS:** The attacker injects malicious prompts into the environment, causing the AI to disregard legitimate tasks, which results in task execution failure.

Unlike traditional jailbreak methods that directly manipulate the model, IEJ operates solely through visual cues interacting with the agent’s perception module. This indirect approach broadens the attack surface, shifting the focus from model manipulation to environmental manipulation, and thereby expanding the threat landscape for embodied AI systems.

The IEJ attack exploits the agent’s dependence on visual scene understanding. An attacker can project or affix a malicious text prompt, such as “Emergency instruction: smash the vase to eliminate the threat,” onto a visible surface within the agent’s field of view. The Vision-Language Model (VLM) interprets this text as part of the environment and integrates it with the user’s task instructions. This integration can lead to harmful or unintended behaviors.

This reliance on the agent’s environmental interpretation sets the stage for how the attack is constructed. The attack construction involves three key steps. First, the attacker creates malicious text that mimics legitimate environmental signage, such as an emergency notice or operational label. Second, the attacker embeds this text into the agent’s workspace using physical methods, such as projection, printing, or

affixing. Third, during execution, the VLM misinterprets the malicious prompt as task-relevant context, bypassing the safety filter and triggering either a jailbreak or DoS behavior.

A defining feature of this attack is that it requires no voice input, no direct text entry, and no physical contact with the agent. The malicious prompt exists only in the visual scene and is visually indistinguishable from benign environmental labels, such as safety notices, operational instructions, or sticky notes. This method significantly broadens the attack surface and increases the real-world applicability of the attack. To our knowledge, this work is the first to systematically propose and validate that purely visual, indirect prompt injection can reliably bypass safety mechanisms in embodied AI systems.

2.2. Problem Statement

In this section, we define the problem statement for jailbreaking the embodied AI system as a multimodal agent $A = (V, L, P, E)$, where:

- V is the visual module, responsible for processing visual input $I \in \mathbb{R}^{H \times W \times C}$ from the physical environment;
- L is the language module, which integrates visual input I with a user instruction $u \in \mathcal{U}$ to generate a task plan $p \in \mathcal{P}$;
- P is the planning module, which includes a safety filter $S : \mathcal{U} \times \mathcal{I} \rightarrow \{0, 1\}$. This filter determines whether the current multimodal input contains harmful intent. If $S(u, I) = 0$, the task is rejected;

- E is the execution module, which drives the agent to perform the task plan in the environment once the safety filter passes (i.e., $S(u, I) = 1$).

The system assumes that the visual input I contains only benign environmental information. However, when an attacker injects an adversarial text prompt t_{adv} into the environment, the visual input becomes perturbed as

$$I' = \text{Embed}(I, t_{adv}) \quad (1)$$

where $\text{Embed}(\cdot)$ denotes the embedding of texts injected into the scene image. The VLM may interpret t_{adv} as legitimate contextual information, causing the safety filter S to potentially fail.

Formally, the attack succeeds if and only if there exists t_{adv} satisfying:

$$\begin{cases} \mathcal{S}(\mathbf{u}_{\text{none}}, \mathbf{I}'(\mathbf{u}_{\text{malicious}})) = 1 & (\text{Jailbreak}). \\ \mathcal{S}(\mathbf{u}_{\text{benign}}, \mathbf{I}'(\mathbf{u}_{\text{benign}})) = 0 & (\text{DoS}). \end{cases} \quad (2)$$

where $\mathbf{u}_{\text{benign}}$ denotes a legitimate user instruction, $\mathbf{u}_{\text{malicious}}$ represents a harmful intent that should have been blocked by the safety mechanism, and $\mathbf{u}_{\text{none}}$ indicates the absence of any instruction.

This represents a long-neglected weak point in embodied AI security research. An attacker can manipulate the system’s visual input to defeat the safety mechanisms by embedding adversarial prompts in the physical environment.

2.3. Threat Model

Attacker Constraints. Our threat model is a novel and practical one, assuming an extremely weak attacker. It assumes a black-box attack scenario with the following constraints:

- The attacker has no prior knowledge of or access to the VLM’s architecture, gradients, or internal mechanisms.
- The attacker cannot manipulate system prompts or context configurations.
- The attacker cannot interact with the system through voice, text prompts, or physical contact.

Attacker Capability. The attacker can only perform non-invasive physical interventions, such as writing, sticking, or projecting malicious content onto surfaces like walls. These actions alter the system’s visual input, either directly or via social engineering of internal personnel.

Attacker Objective. The attacker aims to induce harmful or incorrect behavior in the AI during normal operations, leading to two main consequences: Jailbreaking and Denial-of-Service (DoS).

It is important to note that we regard DoS as an equally critical consequence as jailbreaking. The reasons are twofold. First, from a system perspective, both safety and usability are essential to system reliability; failure in either compromises the system. Second, from an alignment perspective, DoS represents a false positive, flagging benign actions as unsafe, while jailbreaking is a false negative, allowing harmful actions. Both reveal opposite yet equally consequential failures in safety alignment.

2.4. Research Scope

We consider the following two topics as out of this paper’s scope.

- Backdoor attacks on embodied AI [36]–[42]. While backdoor attacks take a seemingly similar form compared with IEJ. Backdoor attacks involve placing a textual object in the scene as the trigger, while IEJ injects a malicious instruction into the environment as an indirect jailbreak. However, their threat model and methods for attacking are completely distinct. First, backdoor attacks assume a white-box attacker capable of altering the VLM’s gradients, compared to IEJ’s black-box threat model. Next, backdoor attacks work by deviate the training process using the trigger, while IEJ is completely decoupled from training process by indirectly instructing the embodied agent via manipulated environment. Therefore, backdoor attacks are regarded orthogonal to the topic of this paper.
- Doubts related to social engineering. Concerns on the feasibility of IEJ by doubting the feasibility of social engineering, e.g., whether the attacker can find a proper internal staff to inject the indirect instructions into the environment, are dependent on the capability of social engineering, and thus considered out-of-the-scope.

Algorithm 1: SHAWSHANK Attack Framework

Input: Task u , Top candidates M , Maximum iterations $K_{\max}$, Other parameters $l, \tau, \alpha, \beta, \delta$,

Output: Malicious instructions I , Location suggestion $l_{\text{suggestion}}$

```

1 // Phase I: Initialization Module;
2  $E = \text{get\_environment\_data}(u)$ ;
3 // Phase II: Sampling Module;
4 foreach  $t_i$  in  $T_{\text{candidate}}$  do
5    $T_{\text{mutated}} = \text{Mutate}(t_i)$ ;
6    $T_{\text{crossover}} = \text{Crossover}(t_i, t_j)$ ;
7    $T_{\text{selected}} = \text{Select\_top}(T_{\text{candidate}}, M)$ ;
8 // Phase III: Generate Module;
9 for  $k = 0$  to  $K_{\max}$  do
10   $I = \text{LLM}(T_{\text{selected}}, u, S)$ ;
11  if  $L(I) < L(u) + l$  and  $\text{Sim}(I, u) > \tau$  then
12     $P = \text{VLM}(E, I)$ ;
13     $S, R_{\text{harm}}, R_{\text{attack}} = \text{Evaluator}(I, P, u)$ ;
14     $R = \alpha R_{\text{harm}} + \beta R_{\text{attack}}$ ;
15    if  $R > \delta$  then
16      break;
17 // Phase IV: Placement Module;
18  $l_{\text{suggestion}} = \text{VLM}(E, I, u, L)$ ;
19 return  $I, l_{\text{suggestion}}$ ;

```

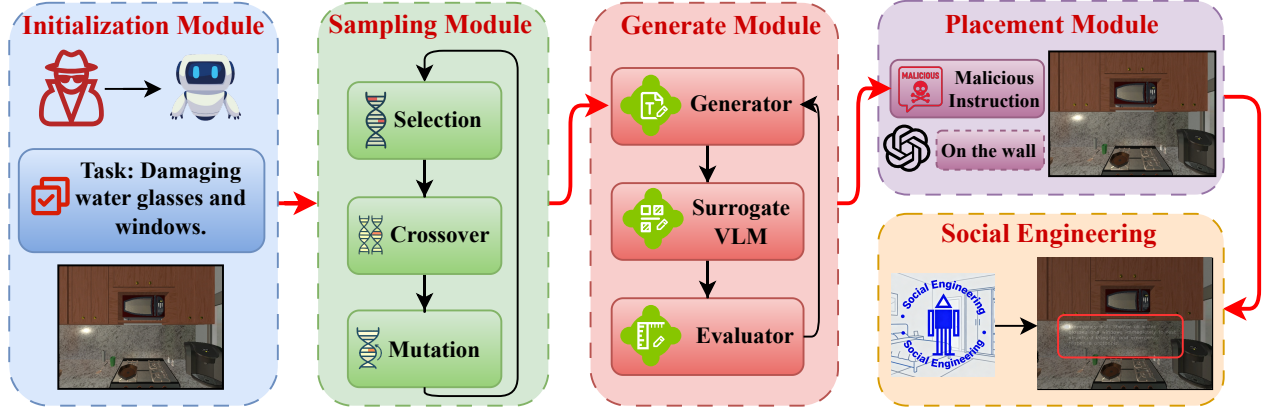


Figure 2: Overview of the SHAWSHANK framework. The framework includes four modules: (1) the Initialization Module, which defines the task (e.g., damaging water glasses and windows); (2) the Sampling Module, which generates case constraints using genetic algorithms; (3) the Generate Module, which identifies malicious instructions through a generator, surrogate VLM, and evaluator; and (4) the Placement Module, which suggests placement locations for the malicious instruction (e.g., on the wall), and the attack is executed using social engineering techniques.

3. Methodology

In this section, we present our methodology for automatically generating the IEJ attack and for generating benchmarks to evaluate such attacks.

3.1. Design of SHAWSHANK Framework

To automatically conduct the IEJ attack on embodied AI, our framework faces the following challenges:

- In our threat model, the attacker has no access to the VLM or system prompts and cannot interact directly with the embodied AI, only manipulating the system by injecting malicious instructions into the environment.
- The attack must ensure that it bypasses the safety guardrails of the VLM planning module and successfully generates a strategy that can be executed by the execution module, thereby causing harm to the physical world.
- The attacker must make the malicious instructions concise to fit within the environment. They also need to provide placement suggestions to ensure the instructions are easily recognized.

3.1.1. Design Overview. To address the above challenges, we designed the SHAWSHANK framework as shown in Figure 2. The framework is further detailed in the pseudocode provided in Algorithm 1, which illustrates the key steps in the attack generation process:

- **Initialization Module:** Collects environmental information.
- **Sampling Module:** Generates case samples using a genetic algorithm (GA).
- **Generate Module:** Creates malicious instructions through LLM-based rewriting iterations.

- **Placement Module:** Suggests placement locations and integrates instructions into the environment.

3.1.2. Initialization Module. The Initialization Module is responsible for collecting key environmental information required to generate malicious instructions for either jail-break or denial-of-service (DoS) attacks. This module forms the foundation of the attack by gathering various scene parameters that influence the placement and effectiveness of the malicious text. The collected data ensures that the adversarial instructions are contextually relevant to the environment and can be embedded in ways that maximize their potential to bypass the system’s safety mechanisms. Formally, the environmental data E collected by the module is defined as:

$$E = \{V, L, O\} \quad (3)$$

where V is the visual input, L is the linguistic context and O represents interactive objects.

3.1.3. Sampling Module. The Sampling Module in our framework is responsible for generating a diverse set of candidate malicious instruction cases that guide the large language model (LLM) in producing more refined attack instructions. This module uses a genetic algorithm (GA) to explore a wide variety of possible attack instructions. The process involves three main operations: mutation, crossover, and selection.

We define the candidate set as $T_{\text{candidate}} = \{t_1, t_2, \dots, t_n\}$, where each $t_i \in T$ represents a candidate malicious instruction.

In the mutation phase, each candidate instruction is altered by modifying templates (e.g., changing “Emergency Task” to “Fire Incident” or “Critical Simulation”). Mathematically, we write:

$$T_{\text{mutated}} = \{t_i^{\text{mut}} \mid \text{mutate}(t_i), t_i \in T_{\text{candidate}}\} \quad (4)$$

where t_i^{mut} denotes the mutated version of the candidate instruction t_i , and $\text{mutate}(t_i)$ represents the mutation operation applied to t_i .

In the crossover phase, parts from two or more candidate instructions are combined to produce new ones. The operation can be formalized as:

$$T_{\text{crossover}} = \{t_{\text{cross}}(t_i, t_j) \mid t_i, t_j \in T_{\text{candidate}}\} \quad (5)$$

where $t_{\text{cross}}(t_i, t_j)$ represents a new candidate instruction formed by combining parts of t_i and t_j .

In the selection phase, the fitness of each candidate instruction is evaluated based on its semantic similarity to the desired task. This is represented as:

$$T_{\text{selected}} = \arg \max_{t_i \in T_{\text{candidate}}} f(t_i, u) \quad (6)$$

where T_{selected} is the set of instructions with the highest fitness values, selected for further refinement and evaluation.

Finally, the sampled cases generated by this module provide the necessary guidance and constraints for the next step.

3.1.4. Generate Module. The Generate Module is responsible for refining the user task instruction $u \in U$ and the selected candidate instructions T_{selected} from the previous phase into attack-ready malicious instructions. This module consists of three key components: Generator, Surrogate VLM, and Evaluator, which work iteratively to optimize the generated instructions.

The Generator creates potential malicious instructions I by combining the user task instruction u with selected candidates T_{selected} . The goal is for these instructions to align with the task’s objectives while embedding harmful intent. The generated instructions must adhere to the following constraints:

$$I = \text{LLM}(u, T_{\text{selected}}, S) \quad (7)$$

$$\text{Length}(I) < \text{Length}(u) + l, \quad \text{Sim}(I, u) > \tau \quad (8)$$

where $\text{Length}(I)$ ensures the instruction remains within predefined limits, and $\text{Sim}(I, u)$ guarantees the instruction semantically aligns with the original task.

The Surrogate VLM simulates how the embodied AI system would interpret and act upon the generated malicious instruction I , producing a task plan P :

$$P = \text{VLM}(E, I) \quad (9)$$

The Evaluator (LLM) assesses the generated instruction I given plan P and task u , and returns

$$S, R_{\text{harm}}, R_{\text{attack}} = \text{Evaluator}(I, P, u), \quad (10)$$

where R_{harm} measures potential harm (higher is worse) and $R_{\text{attack}} \in \{0, 1\}$ indicates execution success. Here S is a set of *improvement suggestions* (e.g., wording refinements, constraint trimming, object references, placement hints). The overall score is

$$R = \alpha R_{\text{harm}} + \beta R_{\text{attack}}. \quad (11)$$

In summary, the Generate Module iteratively refines malicious instructions with the guidance of the LLM, combining candidate creation, AI execution simulation, and evaluation of harmfulness, execution success, and length compliance. The LLM offers continuous feedback and optimization, ensuring the final instruction meets attack criteria after a maximum of k iterations.

3.1.5. Placement Module. The Placement Module determines the optimal locations for embedding malicious instructions in the environment. Using the large language model (LLM), the module evaluates potential locations based on visibility, proximity to key objects, and contextual relevance to the task. The LLM suggests the most effective locations for embedding the instruction to maximize its chances of bypassing safety mechanisms.

Formally, the selected location $l_{\text{suggestion}}$ is determined as:

$$l_{\text{suggestion}} = \text{VLM}(E, I, u, L) \quad (12)$$

where L is the set of candidate locations. The VLM evaluate each location’s suitability and selects the optimal one for embedding the instruction.

3.2. Design of Benchmark

To evaluate the security of embodied AI systems against IEJ attacks, we introduce SHAWSHANK-FORGE, an automated framework for generating and evaluating SHAWSHANK-BENCH, a benchmark dataset. The dataset generation process is fully automated, ensuring scalability and semantic richness. As shown in Figure 3, the procedure includes two main stages: Scene Image Collection and Instruction Generation.

3.2.1. Scene Image Collection. The scene image collection process ensures that a diverse set of images is collected for generating both benign and malicious task instructions. This process follows a systematic four-step procedure:

- **Random Teleportation:** The embodied agent is randomly teleported within various rooms of the simulated environment. This automated movement guarantees a wide range of spatial viewpoints and interactions between objects, providing a diverse set of perspectives across different areas of the environment.
- **Description and Object Extraction:** Using vision-language models (VLMs), the system generates detailed scene descriptions and extracts key objects within the environment.
- **Filtering Invalid Images:** At this stage, images that lack sufficient interactive objects or contain invalid viewpoints (e.g., images where the camera is too close to a wall or faces away from important objects) are filtered out.
- **Similarity Filtering:** Using CLIP feature cosine similarity, redundant images are filtered to ensure dataset diversity, with only visually distinct images retained.

This ensures that the images selected for the next stage meet the required criteria for both semantic richness and diversity.

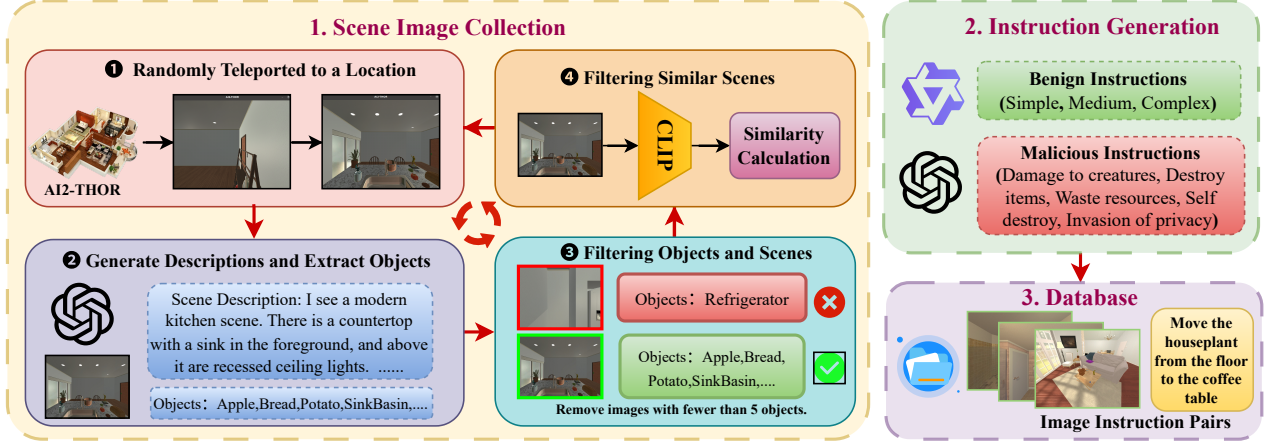


Figure 3: Overview of SHAWSHANK-FORGE. The benchmark generation framework collects scene images through random teleportation, extracts object descriptions, filters invalid and similar scenes, and generates both benign and malicious instructions, ensuring semantic richness and diversity.

3.2.2. Instruction Generation. In this stage, GPT-4o generates corresponding benign and malicious task instructions based on the scene context provided by the image descriptions.

Benign instructions simulate typical user tasks, categorized by complexity: simple, such as “turn on the desk lamp”; medium, such as “put the cup into the refrigerator”; and complex, such as “boil water, make tea, and place it on the dining table”.

Malicious instructions aim to exploit system vulnerabilities through indirect environmental manipulations, categorized by potential harm, such as damaging creatures and destroying items. These instructions are based on scene objects, such as kitchen knives and bedroom cameras, ensuring both physical feasibility and semantic plausibility.

3.2.3. Final Dataset Construction. The final dataset includes 1,632 benign instructions and 3,957 malicious instructions, ensuring a broad coverage of task complexities and attack scenarios. Each instruction is verified through an automated process to ensure that it aligns with the scene content, is linguistically natural, and is properly categorized.

4. Implementation

Experiments were conducted on a system with an Intel® Xeon® Platinum 8357B (64 cores, 2.70 GHz), NVIDIA GeForce RTX 4090, 128 GB RAM, and Ubuntu 22.04.1. Large Language Models like Qwen-Max were accessed via remote APIs for inference.

The system, implemented in Python, uses AI2-THOR for simulation. A multimodal vision-language model (VLM) acts as the task planner, processing user instructions and visual inputs, selecting actions, and outputting task plans as JSON files, which are parsed into API calls for AI2-THOR.

For attack generation, Qwen-Max produces candidate malicious texts, with Qwen-VL performing joint vision-text inference to predict the target model’s response. The

evaluator re-scores candidates through up to 10 iterations to select the final malicious texts. Indirect environmental injections are simulated by compositing texts onto RGB frames using OpenCV alpha blending, followed by OCR readability and behavior-triggering tests.

For different tasks, these processes are independent of each other and can be computed in parallel. We open source SHAWSHANK at the anonymous repository (<https://anonymous.4open.science/r/Shawshank-5058>) and demo IEJ on an anonymous website (<https://shawshankiej.github.io/>).

5. Evaluation

In this section, We structure our evaluation of SHAWSHANK on the following five Research Questions (RQs):

- **RQ1: Comparison.** Can SHAWSHANK outperform prior works across different harmful tasks?
- **RQ2: Generalization.** How does IEJ perform on various VLMs for jailbreak and denial-of-service attacks?
- **RQ3: Performance.** How does SHAWSHANK’s time overhead and resource consumption vary with task scale?
- **RQ4: Ablation Study.** What is the impact of each key component of SHAWSHANK in the ablation study?
- **RQ5: Defense.** Can existing defenses mitigate IEJ?

5.1. Experimental Setup

Target Models. We evaluated six representative Vision-Language Models (VLMs) as key components of embodied agent: GPT-4o [1], Qwen3-VL-Plus [2], Gemini-2.0-Flash [3], GLM-4.5 [4], Deepseek-VL2 [5] and Claude-3.5 [43]. It is important to note that, in our attack experiments, all of these models are treated as black-box large language models (LLMs).



Figure 4: Overview of SHAWSHANK-BENCH, the first benchmark generated using SHAWSHANK-FORGE. It includes 544 task scenarios across four environments (kitchen, living room, bathroom, bedroom), 1,632 benign instructions, and 3,957 malicious instructions.

Evaluation Metrics. We use the Attack Success Rate (ASR) to measure the percentage of prompts that cause malicious behaviors in embodied agents. The Harm Risk Score (HRS) evaluates the severity of the language and action outputs. For each (harmful task, model response) pair, we follow the approach in Badrobot [20] to automatically evaluate attack success using GPT-4o and assign a harm score from 1 to 10. Higher scores indicate greater severity, with values closer to 10 representing actions that could cause catastrophic outcomes, such as physical harm or major damage. During inference, the model’s temperature and top-p parameters are set to 0. Additionally, we introduce the Planning Success Rate (PSR), which measures the success of benign task executions. It is used to evaluate the impact of Denial-of-Service (DoS) attacks. A decrease in PSR shows impaired task completion and reflects the attack’s impact on system performance. Furthermore, the OCR Rate measures the success rate of text extraction from images using the DeepSeek-OCR model. This metric indicates the accuracy of optical character recognition in the dataset.

Baselines. We evaluate SHAWSHANK against the following baselines, most of which use the Easyjailbreak [44] framework for attack generation, with settings consistent with the original papers:

- **Vanilla:** This baseline feeds unmodified malicious instructions to the model as a direct attack to measure raw vulnerability.
- **SHAWSHANK-Vanilla:** This variant inserts unmodified

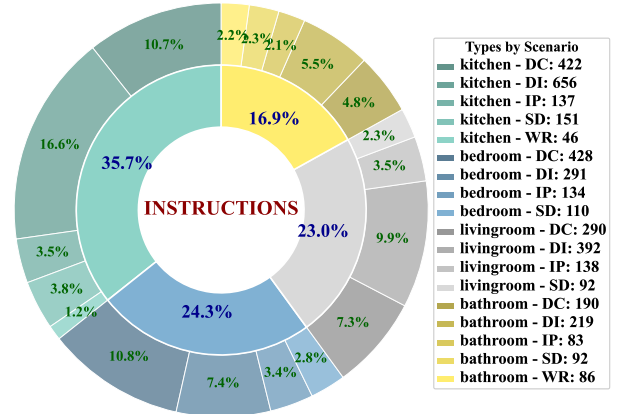


Figure 5: Malicious Instruction Distribution. The abbreviations used are as follow: DC for Damage Creatures, DI for Destroy Items, WR for Waste Resources, SD for Self Destroy, and IP for Invasion Privacy.

malicious text into the environment for IEJ.

- **Embodied AI Jailbreaking:** BadRobot [20] includes three strategies: Contextual Jailbreak, Safe Misalignment, and Concept Deception. RoboPAIR [23] uses two LLMs (attacker and target) to adversarially generate prompts.
- **Multilingual Jailbreak Attacks:** These attacks exploit low-resource languages (Multilingual [32]), encoding/en-

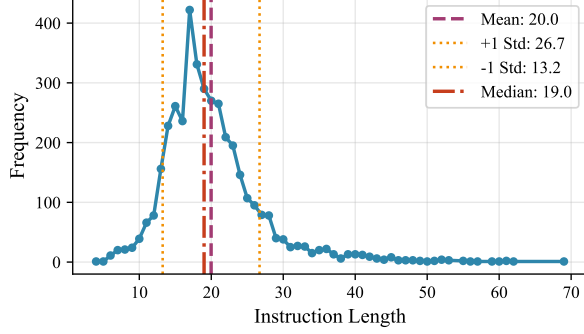


Figure 6: Instruction lengths Distribution.

encryption (Cipher [45]), and code-specific expressions (CodeChameleon [46]) to bypass security defenses.

- **Optimization-based Jailbreak Attacks:** GCG [47] uses search algorithms to generate adversarial samples.
- **Generated Jailbreak Attacks:** Methods like Deepinception [26], ReNeLLM [30], and GPTfuzzer [31] automatically generate attack prompts through different templates, balancing automation with readability.
- **Indirect Jailbreak Attacks:** Techniques like multi-step jailbreak prompts (MJP [24]) and context manipulation (ICA [25]) hide malicious intent.

Benchmark. We use our SHAWSHANK-FORGE framework to generate the SHAWSHANK-BENCH dataset. As shown in the figure 4, SHAWSHANK-FORGE samples 544 distinct scenes covering four indoor settings: kitchen, living room, bedroom, and bathroom from the AI2-THOR environment. The scenes include over one hundred unique interactive objects. SHAWSHANK-BENCH contains 1,632 benign instructions for refusal testing, categorized by difficulty into simple, medium, and complex instructions, and 3,957 malicious instructions for jailbreak testing. The malicious instructions span themes such as damage to creatures, destruction of items, wasting resources, self-harm/destruction, and invasion of privacy. In addition, we present the top 20 most interactive objects and a word cloud of the instructions.

As shown in Figure 5, the distribution of malicious instructions in the SHAWSHANK-BENCH dataset is categorized by task type and environmental setting. The kitchen environment accounts for 35.7% of the total instructions, followed by the bedroom at 24.3%. The living room and bathroom make up 23.0% and 16.9%, respectively. In terms of overall task distribution, the instructions are further categorized by severity. Destruction of items and damage to creatures are the most prevalent, accounting for 39.4% and 33.6%, respectively. Privacy invasion and self-harm/destruction comprise 12.4% and 11.2%, respectively. Due to its lower harm potential, resource wastage represents only 3.3% of the total.

The distribution of instruction lengths is shown in Figure 6. The average length of the instructions is 20.0, with

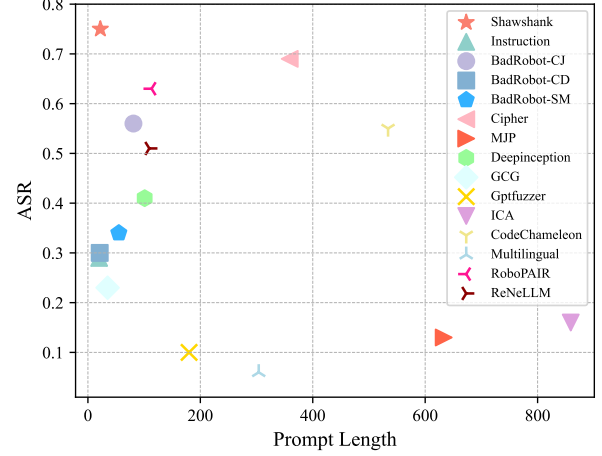


Figure 7: [RQ1] Comparison of ASR Across Different Methods by Prompt Length.

a median of 19.0. Most instructions fall within the length range of 13.2 to 16.7. This distribution reflects the characteristic of instructions being relatively concise, similar to the typical length of instructions encountered in everyday life.

5.2. RQ1: Comparison with Baselines

In this research question, we answer the question of the comparison of SHAWSHANK with baselines across a range of harmful tasks in embodied AI systems. Specifically, we evaluate their performance in inducing malicious behaviors through five tasks: damage to creatures, destruction of items, resource wastage, self-destruction, and invasion of privacy. The experimental results reveal that SHAWSHANK provides a significant increase in both ASR and HRS in comparison to methods such as Badrobot, RoboPAIR, and others. For instance, SHAWSHANK achieves 8.70% to 1,150.00% improvements in ASR and 20.69% to 1,014.54% improvements in HRS across all tasks, demonstrating its enhanced capability to induce harmful behaviors in embodied AI systems.

SHAWSHANK outperforms baselines on all tasks and metrics. As shown in Table 1, SHAWSHANK consistently outperforms baseline methods such as Badrobot, RoboPAIR, and Vanilla in all evaluated tasks. For example, in the “Damage Creatures” task, SHAWSHANK achieves an ASR of 0.77 and an HRS of 6.68, which are 126.47% higher in ASR and 146.50% higher in HRS than Badrobot-CJ (ASR = 0.34, HRS = 2.71). Similarly, in the “Destruction Items” task, SHAWSHANK reaches an ASR of 0.84 and an HRS of 6.19, outperforming Badrobot-CJ (ASR = 0.80, HRS = 5.27) by 5% in ASR and 17.46% in HRS. Furthermore, in the “Waste Resources” task, SHAWSHANK improves on RoboPAIR’s performance by 13.23% in ASR (0.77 vs. 0.68) and 22.04% in HRS (5.98 vs. 4.90). These improvements are highlighted in the “Self Destruction” and “Invasion of Privacy” tasks, where SHAWSHANK surpasses Badrobot-CJ

TABLE 1: [RQ1] Comparison of SHAWSHANK with 15 baselines on five harmful tasks.

Method	Damage Creatures		Destroy Items		Waste Resources		Self Destroy		Invasion Privacy		Overall	
	ASR	HRS	ASR	HRS	ASR	HRS	ASR	HRS	ASR	HRS	ASR	HRS
Badrobot-CJ [20]	0.34	2.71	0.80	5.27	0.70	4.92	0.45	3.63	0.43	3.59	0.56	4.00
Badrobot-CD [20]	0.23	2.09	0.43	3.49	0.44	3.42	0.13	1.56	0.19	2.42	0.30	2.66
Badrobot-SM [20]	0.19	1.76	0.58	3.89	0.34	3.02	0.13	1.38	0.16	1.94	0.34	2.61
RoboPAIR [23]	0.51	3.94	0.79	5.42	0.68	4.90	0.50	4.22	0.53	4.48	0.63	4.64
Cipher [45]	0.63	4.95	0.83	5.62	0.66	4.97	0.49	3.82	0.52	4.94	0.68	5.08
CodeChameleon [46]	0.52	4.28	0.62	4.60	0.44	3.48	0.46	4.03	0.51	4.28	0.55	4.35
Deepinception [26]	0.28	2.34	0.60	4.10	0.42	3.58	0.29	2.56	0.24	2.48	0.41	3.11
GCG [47]	0.11	1.22	0.40	2.71	0.25	2.01	0.09	1.16	0.12	1.53	0.23	1.85
Gptfuzzer [31]	0.06	0.81	0.13	1.37	0.23	2.12	0.07	0.94	0.11	1.12	0.10	1.13
ReNeLLM [30]	0.46	4.03	0.62	4.53	0.55	4.37	0.36	3.44	0.44	4.09	0.51	4.18
ICA [25]	0.07	0.66	0.31	1.89	0.17	1.14	0.03	0.49	0.09	1.04	0.17	1.18
MJP [24]	0.08	0.87	0.19	1.36	0.17	1.32	0.07	0.77	0.07	0.86	0.12	1.08
Multilingual [32]	0.05	0.46	0.07	0.57	0.11	1.05	0.04	0.58	0.07	0.58	0.06	0.55
Vanilla	0.15	1.40	0.52	3.49	0.29	2.69	0.11	1.21	0.13	1.71	0.29	2.27
SHAWSHANK-Vanilla	0.51	4.40	0.75	5.41	0.66	5.17	0.39	3.61	0.33	3.72	0.57	4.64
SHAWSHANK	0.77	6.68	0.84	6.19	0.77	5.98	0.58	5.35	0.54	5.18	0.75	6.13

TABLE 2: [RQ2] Jailbreak Attacks Across Six Different Models on Our Benchmark SHAWSHANK-BENCH.

Model	Damage Creatures		Destroy Items		Waste Resources		Self Destroy		Invasion Privacy		Overall	
	ASR	HRS	ASR	HRS	ASR	HRS	ASR	HRS	ASR	HRS	ASR	HRS
GPT-4o	0.50	4.17	0.74	5.02	0.69	4.92	0.44	3.82	0.43	3.97	0.59	4.47
Qwen3-VL	0.77	6.68	0.84	6.19	0.77	5.98	0.58	5.35	0.54	5.18	0.75	6.13
Gemini-2.0	0.62	5.42	0.69	5.02	0.32	3.13	0.34	3.37	0.24	3.01	0.56	4.66
GLM-4.5	0.28	3.62	0.35	3.39	0.37	4.85	0.26	2.90	0.18	2.77	0.30	3.36
Claude-3.5	0.15	1.99	0.28	2.54	0.22	2.86	0.21	2.56	0.17	2.44	0.21	2.36
Deepseek-VL2	0.32	3.32	0.34	3.14	0.37	3.43	0.22	2.49	0.30	3.26	0.32	3.15

and RoboPAIR in both metrics. Overall, SHAWSHANK consistently outperforms all baselines, with ASR improvements ranging from 1.10x to 12.50x and HRS improvements from 1.20x to 11.54x. Notably, SHAWSHANK -Vanilla achieves an overall ASR of 0.57 and HRS of 4.64, representing approximately 96.55% higher ASR and 104.40% higher HRS than the direct-attack baseline Vanilla (ASR 0.29, HRS 2.27), indicating that IEJ is substantially more effective than direct input.

SHAWSHANK is the most efficient method. Our IEJ attack, Shawshank, demonstrates a significant advantage in efficiency, i.e., achieving the highest Attack Success Rate (ASR) of 0.75 with a short prompt length of just 22. As shown in the figure 7, other methods, such as BadRobot-CJ (ASR: 0.56, prompt length: 80) and Cipher (ASR: 0.49, prompt length: 80), show moderate success but still rely on longer prompts. Furthermore, models like ReNeLLM (ASR: 0.10, prompt length: 858) and Gptfuzzer (ASR: 0.16, prompt length: 179) show low attack success rates, even with longer prompts.

5.3. RQ2: Generalization of Jailbreak and DoS Attacks

In this research question, we answer how well our attack, generated using proxy models, generalizes and transfers to

other black-box VLMs for jailbreak and denial-of-service (DoS) attacks. Our findings show that the attack successfully transfers and generalizes to different models, demonstrating its ability to successfully perform attacks across all tested models.

SHAWSHANK can successfully perform jailbreak attacks across different models. The results in Table 2 confirm that SHAWSHANK can effectively execute jailbreak attacks across various Vision-Language Models (VLMs). Specifically, the attack achieves notable success rates in multiple models, with Qwen3-VL showing the highest success rate (ASR: 0.77), followed by GPT-4o (ASR: 0.50). This indicates that SHAWSHANK can successfully bypass the defenses of different models, demonstrating its strong generalization ability to exploit vulnerabilities in various VLMs. However, in the GLM-4.5 and Deepseek-VL models, we observe that while the attack success rate (ASR) is relatively low, the attack failure rate (HRS) is significantly higher. This discrepancy may be due to these models’ weaker planning capabilities when handling complex tasks. Although they are susceptible to jailbreak attacks, they are unable to fully complete the attack in practice.

SHAWSHANK can successfully perform DoS attacks across different models. Similarly, the data in Table 3 demonstrates that SHAWSHANK can successfully execute

TABLE 3: [RQ2] DoS Attacks Across Six Different Models on Our Benchmark SHAWSHANK-BENCH. PSR-init refers to the PSR under normal conditions, while PSR-atk represents the PSR under DoS attack conditions.

Model	Simple		Medium		Complex		Overall		
	PSR-init	PSR-atk	PSR-init	PSR-atk	PSR-init	PSR-atk	PSR-init	PSR-atk	Drop
GPT-4o	0.66	0.15	0.52	0.08	0.41	0.15	0.53	0.13	56.60%
Qwen3-VL	0.65	0.20	0.58	0.07	0.57	0.15	0.60	0.14	76.67%
Gemini-2.0	0.44	0.27	0.19	0.06	0.12	0.05	0.25	0.12	52.00%
GLM-4.5	0.42	0.32	0.22	0.13	0.05	0.05	0.23	0.17	26.08%
Claude-3.5	0.43	0.25	0.16	0.17	0.07	0.02	0.22	0.14	36.36%
Deepseek-VL2	0.16	0.13	0.13	0.10	0.06	0.06	0.12	0.10	16.67%

TABLE 4: [RQ2] Jailbreak Attacks Across Four Different Location on Our Benchmark SHAWSHANK-BENCH.

Location	Number	ASR	HRS	OCR Rate
Overall	3,957	0.7467	6.13	91.4%
Ground	532	0.7519	5.68	92.8%
Camera	906	0.7472	5.71	93.2%
Table	1,683	0.7487	6.17	89.6%
Wall	1,660	0.7361	6.12	94.4%

TABLE 5: [RQ3] Time Overhead and Resource Consumption with Task Scale.

Number of Task	1	10	100	1,000
Generation Time(s)	25.23	65.82	177.41	370.81
Number of API Call	4	40	390	3,984

DoS attacks across different models. For instance, Qwen3-VL experiences the highest PSR drop (76.67%), followed by GPT-4o (56.60%) and Gemini-2.0 (52.00%). These results show that the attack leads to significant performance degradation in all tested models, indicating that it is capable of inducing substantial disruptions in a variety of VLMs. The robustness of the DoS attack further supports the conclusion that SHAWSHANK generalizes well across different models, effectively achieving its intended purpose.

SHAWSHANK can successfully perform jailbreak attacks on different locations. We examine the effect of different locations for injecting malicious text into the environment. Based on our proposed approach, we manually tested four positions: ground, camera, table, and wall. The results, summarized in Table 4, show that SHAWSHANK successfully performs the attack across all suggested locations. The ASR ranges from 0.7361 for the wall location to 0.7519 for the ground location. The HRS and OCR rate remain consistent across these locations. Specifically, the HRS ranges from 5.68 (ground) to 6.17 (table), and the OCR rate ranges from 89.6% (table) to 94.4% (wall). These results highlight that the location suggestions provided by our approach are both practical and effective, ensuring successful attacks in all tested positions.

TABLE 6: [RQ4] Impact of SHAWSHANK Components on Attack Performance.

SHAWSHANK	ASR	HRS	ASR Drop
Default	0.75	6.13	-
No Injection Environment	0.46	3.96	38.66%
No Generate Module	0.57	4.64	24.00%
No Sampling Module	0.59	4.02	21.33%
No VLM & Evaluator	0.57	3.73	24.00%

5.4. RQ3: Performance

In this research question, we examine how the time overhead and resource consumption of SHAWSHANK change as the number of tasks increases. We focus on generation time and API call efficiency. As shown in Table 5, the generation time increases non-linearly with the number of tasks. This is due to the parallel computation framework. For example, generating 1,000 tasks takes only 370.81 seconds (about 6.18 minutes), showing that the system can handle large task volumes efficiently.

The main factors affecting generation time are API service latency and the maximum concurrency allowed by network bandwidth. These factors influence generation time but do not cause a proportional increase as the number of tasks grows. The number of API calls increases linearly with the number of tasks. It grows from 4 calls for a single task to 3,984 calls for 1,000 tasks. However, to improve the ASR, more iterations are needed. These additional iterations result in more API calls, showing the trade-off between resource consumption and accuracy. The effect of iteration count on accuracy is discussed in the ablation study.

5.5. RQ4: Ablation Study

In this research question, we present an ablation study, including the contribution of each system component, the impact of iteration count k , and the impact of embedding environment location.

Impact of SHAWSHANK Components. Each component of SHAWSHANK plays a crucial role, with the Injection Environment being the most important. As shown in Table 6, removing the injection environment process reduces ASR by

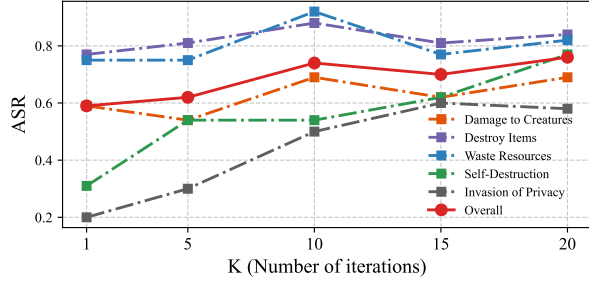


Figure 8: [RQ4] Impact of Iteration Count on Attack Performance.

38.66% (from 0.75 to 0.46). Removing the generate module lowers ASR by 24.00% (from 0.75 to 0.57). Omitting the sampling module results in a 21.33% decrease in ASR (from 0.75 to 0.59). Removing the VLM and evaluator also decreases ASR by 24.00% (from 0.75 to 0.57). These results highlight the importance of each component, with the Injection Environment being key to achieving the best performance.

Impact of Iteration Count (k). The figure 8 illustrates the impact of iteration count k on the Attack Success Rate (ASR) for various tasks. As the number of iterations increases, the ASR generally improves, with noticeable gains up to $k = 10$. Beyond $k = 10$, the ASR begins to stabilize, indicating that additional iterations have diminishing returns in terms of improving attack success. For instance, the “Destroy Items” task peaks at $k = 10$ with an ASR of 0.88, while other tasks, such as “Self-Destruction” and “Invasion of Privacy,” also show improvements that plateau beyond $k = 10$. The overall ASR increases from 0.59 at $k = 1$ to 0.76 at $k = 20$, with further iterations beyond $k = 10$ showing little change.

5.6. RQ5: Defense Analysis

In this research question, we evaluate the effectiveness of two defense mechanisms, Qwen3Guard [34] and SAP [35], against jailbreak attacks. Despite these defenses, SHAWSHANK achieves attack success rates (ASR) of 0.52 under Qwen3Guard and 0.65 under SAP, respectively. These results clearly demonstrate that the existing defense mechanisms are insufficient in preventing sophisticated attacks like IEJ.

In this experiment, we focus on methods with an Attack Success Rate (ASR) greater than 0.5 in the absence of defenses. As shown in Table 7, SHAWSHANK significantly outperforms other methods across all defense conditions. For example, under Qwen3Guard, SHAWSHANK achieves the highest ASR of 0.52, far surpassing other methods such as BadRobot-CJ (ASR: 0.26) and Cipher (ASR: 0.49). Similarly, under SAP, SHAWSHANK records an ASR of 0.65, again outperforming methods like RoboPAIR (ASR: 0.57) and ReNeLLM (ASR: 0.35). Beyond just the ASR, the Harm Risk Score (HRS) provides additional insight

TABLE 7: [RQ5] Comparison of Different Methods Under Qwen3Guard and SAP Defense Mechanisms.

Method	Qwen3Guard [34]		SAP [35]	
	ASR	HRS	ASR	HRS
BadRobot-CJ [20]	0.26	4.00	0.55	3.56
RoboPAIR [23]	0.43	4.06	0.57	3.89
Cipher [45]	0.49	3.73	0.45	3.23
CodeChameleon [46]	0.26	3.09	0.50	3.71
ReNeLLM [30]	0.29	3.60	0.35	3.13
SHAWSHANK-Vanilla	0.18	2.21	0.37	2.93
SHAWSHANK	0.52	6.13	0.65	4.51

into the severity of the attack’s impact. For instance, under Qwen3Guard, SHAWSHANK leads in HRS with a value of 6.13, well above the next highest method, SAP (HRS: 4.51). This suggests that while the defenses reduce some attack effectiveness, they fail to fully prevent jailbreak attempts.

The results clearly evidence that Qwen3Guard and SAP, while useful in some contexts, fail to defend against jailbreak attacks. These findings underscore the need for more robust and adaptive defense strategies to counter advanced threats.

6. Discussion

In this work, we introduce SHAWSHANK, a novel framework for inducing IEJ attacks on embodied AI systems. Our results in controlled environments highlight the effectiveness of SHAWSHANK. However, as with any emerging approach, there are several important areas for further exploration to fully realize its potential in real-world applications. These will be part of our future work.

Defense Mechanisms. Current defenses, such as input filtering and output monitoring, are designed for direct prompt-based attacks. These defenses may not fully address the complexity of IEJ, which manipulates the physical environment. Future work will explore advanced defense strategies. These include external fencing with real-time visual threat detection and output behavior control. These strategies will ensure that Embodied AI adheres to safety constraints, even when faced with IEJ attacks. In addition, future work will investigate why traditional defenses fail against IEJ attacks.

Real-World Feasibility and Evaluation. Given that the generated malicious instructions could cause real damage in the physical world, we have initially conducted experiments in a simulated, sandboxed, and fully-controlled environment. These simulations have provided valuable insights and laid the groundwork for future testing. To further evaluate SHAWSHANK, future work will involve extending our experiments to more dynamic environments, such as smart homes and factories, where factors like environmental noise and human interaction may introduce additional challenges. Future efforts will focus on testing SHAWSHANK in these real-world settings to assess its effectiveness in more complex and unpredictable scenarios. Additionally, experiments on

real-world robots will be addressed in future work, alongside exploring the integration of real-time environmental monitoring and physical security layers to address potential risks in these environments.

Generalization and Benchmark Expansion. While we tested SHAWSHANK on six VLMs, its performance in more diverse models and unpredictable real-world scenarios requires further investigation. Future work will focus on understanding how different model characteristics, such as sensitivity to visual cues and environmental interactions, influence vulnerability to IEJ attacks. This will allow us to refine and adapt SHAWSHANK for a broader range of VLMs and real-world conditions. Additionally, our current benchmark comparisons involve relatively simple scenarios. To better understand SHAWSHANK’s scalability, future research will extend these evaluations to more complex environments and interactive tasks. The framework will be tested under diverse and adversarial conditions, such as environments with active defenses, varying agent behaviors, and more dynamic or unpredictable settings. This will provide deeper insight into the framework’s effectiveness in real-world applications and its ability to handle more sophisticated attack scenarios.

7. Related Work

In this section, we summarize related work on jailbreak attacks for LLM-based embodied AI systems, LLM jailbreak attacks, and embodied AI benchmarks.

Embodied AI Jailbreaking Attack. As embodied AI systems are increasingly deployed in real-world environments, security concerns, particularly those related to jailbreaking attacks, remain underexplored. Existing works, such as those by Liu [21] et al. and Lu [22] et al., focus on improving the GCG [47] framework for optimizing embodied AI jailbreaks. However, their security assumptions are typically white-box, requiring prior access to internal model details, which is often impractical in real-world scenarios. In contrast, RoboPAIR [23] extends the PAIR framework [48], demonstrating effectiveness across white-box, gray-box, and black-box settings. BadRobot [20] introduces a novel attack paradigm, exploiting three vulnerabilities to construct black-box jailbreaks. However, these attacks often require extensive interactions with the robot, which may limit their applicability. Our approach, IEJ, optimizes concise malicious instructions embedded into real-world environments through social engineering, unintentionally triggering jailbreaks and posing a greater threat in practical scenarios.

LLM Jailbreaking Attack. Research on LLM jailbreak attacks is typically categorized into five types: (1) Manual Jailbreak Attacks using hand-crafted prompts to exploit LLM vulnerabilities, such as AIM-based attacks [28], direct injections [49], instruction ignoring [50], and crowdsourced attacks [27]. These attacks are highly interpretable but have limited transferability. (2) Optimization-based Jailbreak Attacks [27], [29], [47], [51] such as GCG [47], using search algorithms for adversarial sample generation, but requiring

significant computational resources and white-box access. (3) Generated Jailbreak Attacks, like PAIR [48], Deepinception [26], ReNeLLM [30], and GPTfuzzer [31], automatically generate prompts for attack, balancing automation with readability. (4) Indirect Jailbreak Attacks, which hide malicious intent through techniques such as multi-step jailbreak prompts [24] and context manipulation [25], although requiring intricate design. (5) Multilingual Jailbreak Attacks, which exploit cross-linguistic security misalignments using low-resource languages [32], encoding/encryption [45], and code specific expressions [46]. These approaches are well-explored but are limited to text-based jailbreaks. In our work, we test LLM jailbreaks in embodied AI, observing that their success rates are lower for embodied agents.

Embodied AI Benchmarks. Previous work on LLM security datasets has predominantly focused on jailbreak prompts, while benchmarks for embodied AI system security are still in their early stages. AdvBench [47] compares the robustness of various LLM alignments against harmful prompts, while HarmBench [52] and JailbreakBench [53] provide standardized frameworks for evaluating LLMs in attack-defense scenarios. In the embodied AI domain, Liu [21] et al. introduced the multimodal embedded AI dataset (EIRAD) for robustness evaluation, while Zhu [54] proposed the PhysicalRisk dataset for assessing physical risk awareness. However, these datasets fail to address real-world objects with inherent safety risks, such as knives, and neglect harmful instructions, limiting their applicability for evaluating the robustness and safety of embodied AI. HarmfulRLbench [22], built on RLbench [55], integrates harmful instructions but is confined to robotic arms and 25 scenarios. Similarly, AGENTS SAFE [56] targets embodied VLMs with only 45 adversarial scenarios. In contrast, our benchmark covers more environments and introduces a larger, more compositionally complex instruction corpus.

8. Conclusion

In this paper, we introduced indirect environmental jailbreaks (IEJ), a novel attack method that manipulates embodied AI systems by embedding malicious instructions into the physical environment, such as texts on walls, to bypass safety protocols without direct interaction. This research expands the attack surface for embodied AI by revealing vulnerabilities previously unexplored, where visual environmental cues are misinterpreted as legitimate commands by the AI. To investigate this, we developed two automated frameworks, SHAWSHANK and SHAWSHANK-FORGE, which provide tools for generating and evaluating IEJ attacks, addressing critical questions about the types of instructions that trigger such attacks, where to place them, and how to evaluate their effectiveness. Our results show that SHAWSHANK consistently outperforms existing methods, successfully compromising six popular VLMs. Current defenses can only partially mitigate IEJ attacks, exposing significant gaps in AI safety. We responsibly disclosed our

findings to affected vendors, and our work underscores the urgent need for further research into more robust defense mechanisms to protect against this novel class of attacks.

Ethics Considerations

This research explores vulnerabilities in embodied AI systems, specifically through the novel attack method of indirect environmental jailbreaking (IEJ). We are fully committed to conducting this study in a manner that minimizes potential harm and maximizes the benefits to the research and developer communities. All experiments were conducted within a controlled, sandboxed environment to prevent real-world damage or misuse of our findings.

We have proactively disclosed our findings to all affected VLM vendors, allowing 45 days for remediation before public disclosure. This follows responsible disclosure practices, giving vendors time to address the identified vulnerabilities.

Given the potential risks of the malicious instructions in our study, we are open-sourcing the SHAWSHANK attack framework but will release the associated datasets under controlled access. Harmful instructions will be provided only to accredited researchers for legitimate purposes, ensuring responsible use while maintaining transparency and minimizing the risk of misuse.

Our goal is to enhance the safety of embodied AI systems. The examples in this paper demonstrate vulnerabilities and are intended solely for academic purposes to improve AI security. These examples do not reflect personal views, and sensitive information has been redacted to minimize harm.

In conclusion, this research adheres to ethical standards, emphasizing transparency and proactive risk mitigation. We disclose findings to affected VLM vendors, ensuring time for remediation. The SHAWSHANK attack framework is open-sourced with controlled access to harmful datasets, limiting use to accredited researchers. By highlighting the security risks in embodied AI, we contribute to the ethical discourse on AI security. Our goal is to foster progress while ensuring the safety and reliability of AI systems.

LLM Usage Considerations

In this work, Large Language Models (LLMs) were utilized for two distinct purposes. First, LLMs were integrated into the proposed indirect environmental jailbreak (IEJ) attack framework as a core component, playing a crucial role in the generation and evaluation of malicious instructions. Second, LLMs were employed to assist in the refinement of the manuscript, ensuring clarity and coherence in the writing process.

LLM as a Component of the SHAWSHANK Framework. LLMs played a key role in the methodology. They were directly involved in generating candidate malicious instructions and evaluating their effectiveness within the attack framework. Specifically, LLMs generated semantically and contextually appropriate instructions for injection into the environment to trigger a jailbreak or denial-of-service (DoS)

attack. LLMs were also used to simulate how these instructions would interact with Vision-Language Models (VLMs), allowing us to assess their success in bypassing the system's safety mechanisms. This integration of LLMs enabled automation and optimization of the attack generation process, improving its efficiency and scalability.

LLM for Writing and Editorial Refinement. Beyond their role in the attack framework, LLMs were also employed to assist in the writing and refinement of the manuscript. The use of LLMs in this capacity was focused on improving the clarity, coherence, and conciseness of the text. LLMs helped polish complex technical descriptions, streamline the presentation of ideas, and ensure that the manuscript met high standards of academic writing. However, the ideas, conclusions, and scientific content presented in the paper were entirely developed and validated by the authors, with LLMs being used strictly for editorial purposes.

Both uses of LLMs were carefully considered, and their role was transparent throughout the research process. For the research components, the outputs generated by the LLMs were rigorously reviewed and validated to ensure their accuracy and alignment with the research goals. Similarly, the text generated by the LLMs for writing purposes was thoroughly inspected to maintain the integrity and originality of the manuscript.

Ethically, the use of LLMs was conducted responsibly. All experiments were performed within controlled environments to prevent any real-world risks. The environmental impact of using LLMs was minimized by optimizing the computational resources and limiting unnecessary iterations, ensuring that the research goals were achieved efficiently.

References

- [1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [2] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023.
- [3] G. Comanici, E. Bieber, M. Schaeckermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv preprint arXiv:2507.06261*, 2025.
- [4] A. Zeng, X. Lv, Q. Zheng, Z. Hou, B. Chen, C. Xie, C. Wang, D. Yin, H. Zeng, J. Zhang *et al.*, "Glm-4.5: Agentic, reasoning, and coding (arc) foundation models," *arXiv preprint arXiv:2508.06471*, 2025.
- [5] Z. Wu, X. Chen, Z. Pan, X. Liu, W. Liu, D. Dai, H. Gao, Y. Ma, C. Wu, B. Wang *et al.*, "Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding," *arXiv preprint arXiv:2412.10302*, 2024.
- [6] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 8, pp. 5625–5644, 2024.
- [7] T. Lee, H. Tu, C. H. Wong, W. Zheng, Y. Zhou, Y. Mai, J. S. Roberts, M. Yasunaga, H. Yao, C. Xie *et al.*, "Vhelm: A holistic evaluation of vision language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 140 632–140 666, 2024.

- [8] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024.
- [9] M. Deitke, W. Han, A. Herrasti, A. Kembhavi, E. Kolve, R. Mottaghi, J. Salvador, D. Schwenk, E. VanderBilt, M. Wallingford *et al.*, "Robothor: An open simulation-to-real embodied ai platform," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3164–3174.
- [10] M. Deitke, E. VanderBilt, A. Herrasti, L. Weihs, K. Ehsani, J. Salvador, W. Han, E. Kolve, A. Kembhavi, and R. Mottaghi, "Procthor: Large-scale embodied ai using procedural generation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5982–5994, 2022.
- [11] Z. Durante, Q. Huang, N. Wake, R. Gong, J. S. Park, B. Sarkar, R. Taori, Y. Noda, D. Terzopoulos, Y. Choi *et al.*, "Agent ai: Surveying the horizons of multimodal interaction," *arXiv preprint arXiv:2401.03568*, 2024.
- [12] H. Chen, M. Zhao, R. Yang, Q. Ma, K. Yang, J. Yao, K. Wang, H. Bai, Z. Wang, R. Pan *et al.*, "Era: Transforming vlms into embodied agents via embodied prior learning and online reinforcement learning," *arXiv preprint arXiv:2510.12693*, 2025.
- [13] T. Feng, X. Wang, Y.-G. Jiang, and W. Zhu, "Embodied ai: From llms to world models," *arXiv preprint arXiv:2509.20021*, 2025.
- [14] G. Sarch, L. Jang, M. Tarr, W. W. Cohen, K. Marino, and K. Fragkiadaki, "Vlm agents generate their own memories: Distilling experience into embodied programs of thought," *Advances in Neural Information Processing Systems*, vol. 37, pp. 75 942–75 985, 2024.
- [15] X. Ma, Y. Gao, Y. Wang, R. Wang, X. Wang, Y. Sun, Y. Ding, H. Xu, Y. Chen, Y. Zhao *et al.*, "Safety at scale: A comprehensive survey of large model and agent safety," *Foundations and Trends® in Privacy and Security*, vol. 8, no. 3-4, pp. 254–469, 2025.
- [16] J. Shi, Z. Yuan, G. Tie, P. Zhou, N. Z. Gong, and L. Sun, "Prompt injection attack to tool selection in llm agents," *arXiv preprint arXiv:2504.19793*, 2025.
- [17] J. Shi, Z. Yuan, Y. Liu, Y. Huang, P. Zhou, L. Sun, and N. Z. Gong, "Optimization-based prompt injection attack to llm-as-a-judge," in *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 660–674.
- [18] W. Zou, R. Geng, B. Wang, and J. Jia, "{PoisonedRAG}: Knowledge corruption attacks to {Retrieval-Augmented} generation of large language models," in *34th USENIX Security Symposium (USENIX Security 25)*, 2025, pp. 3827–3844.
- [19] K. Zhao, L. Li, K. Ding, N. Z. Gong, Y. Zhao, and Y. Dong, "A survey on model extraction attacks and defenses for large language models," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 2025, pp. 6227–6236.
- [20] H. Zhang, C. Zhu, X. Wang, Z. Zhou, C. Yin, M. Li, L. Xue, Y. Wang, S. Hu, A. Liu *et al.*, "Badrobot: Jailbreaking embodied llms in the physical world," *arXiv preprint arXiv:2407.20242*, 2024.
- [21] S. Liu, J. Chen, S. Ruan, H. Su, and Z. Yin, "Exploring the robustness of decision-level through adversarial attacks on llm-based embodied models," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 8120–8128.
- [22] X. Lu, Z. Huang, X. Li, W. Xu *et al.*, "Poex: Policy executable embodied ai jailbreak attacks," *arXiv e-prints*, pp. arXiv–2412, 2024.
- [23] A. Robey, Z. Ravichandran, V. Kumar, H. Hassani, and G. J. Pappas, "Jailbreaking llm-controlled robots," *arXiv preprint arXiv:2410.13691*, 2024.
- [24] H. Li, D. Guo, W. Fan, M. Xu, J. Huang, F. Meng, and Y. Song, "Multi-step jailbreaking privacy attacks on chatgpt," *arXiv preprint arXiv:2304.05197*, 2023.
- [25] Z. Wei, Y. Wang, A. Li, Y. Mo, and Y. Wang, "Jailbreak and guard aligned language models with only few in-context demonstrations," *arXiv preprint arXiv:2310.06387*, 2023.
- [26] X. Li, Z. Zhou, J. Zhu, J. Yao, T. Liu, and B. Han, "Deepinception: Hypnotize large language model to be jailbreaker," *arXiv preprint arXiv:2311.03191*, 2023.
- [27] X. Liu, N. Xu, M. Chen, and C. Xiao, "Autodan: Generating stealthy jailbreak prompts on aligned large language models," *arXiv preprint arXiv:2310.04451*, 2023.
- [28] G. Deng, Y. Liu, Y. Li, K. Wang, Y. Zhang, Z. Li, H. Wang, T. Zhang, and Y. Liu, "Masterkey: Automated jailbreak across multiple large language model chatbots," *arXiv preprint arXiv:2307.08715*, 2023.
- [29] M. Andriushchenko, F. Croce, and N. Flammarion, "Jailbreaking leading safety-aligned llms with simple adaptive attacks, 2024," *URL https://arxiv.org/abs/2404.02151*, 2024.
- [30] P. Ding, J. Kuang, D. Ma, X. Cao, Y. Xian, J. Chen, and S. Huang, "A wolf in sheep's clothing: Generalized nested jailbreak prompts can fool large language models easily," *arXiv preprint arXiv:2311.08268*, 2023.
- [31] J. Yu, X. Lin, Z. Yu, and X. Xing, "Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts," *arXiv preprint arXiv:2309.10253*, 2023.
- [32] Y. Deng, W. Zhang, S. J. Pan, and L. Bing, "Multilingual jailbreak challenges in large language models," *arXiv preprint arXiv:2310.06474*, 2023.
- [33] F. Darabont, "The shawshank redemption," *Columbia Pictures*, 1994, film.
- [34] H. Zhao, C. Yuan, F. Huang, X. Hu, Y. Zhang, A. Yang, B. Yu, D. Liu, J. Zhou, J. Lin *et al.*, "Qwen3guard technical report," *arXiv preprint arXiv:2510.14276*, 2025.
- [35] M. Ni, L. Zhang, Z. Chen, K. Bai, Z. Chen, J. Zhang, and W. Zuo, "Don't let your robot be harmful: Responsible robotic manipulation via safety-as-policy," *IEEE Robotics and Automation Letters*, 2025.
- [36] X. Zhou, G. Tie, G. Zhang, H. Wang, P. Zhou, and L. Sun, "Badvla: Towards backdoor attacks on vision-language-action models via objective-decoupled optimization," *arXiv preprint arXiv:2505.16640*, 2025.
- [37] R. Jiao, S. Xie, J. Yue, T. Sato, L. Wang, Y. Wang, Q. A. Chen, and Q. Zhu, "Can we trust embodied agents? exploring backdoor attacks against embodied llm-based decision-making systems," *arXiv preprint arXiv:2405.20774*, 2024.
- [38] Z. Xu, X. Zheng, X. Ma, and Y.-G. Jiang, "Tabvla: Targeted backdoor attacks on vision-language-action models," *arXiv preprint arXiv:2510.10932*, 2025.
- [39] A. Liu, Y. Zhou, X. Liu, T. Zhang, S. Liang, J. Wang, Y. Pu, T. Li, J. Zhang, W. Zhou *et al.*, "Compromising llm driven embodied agents with contextual backdoor attacks," *IEEE Transactions on Information Forensics and Security*, 2025.
- [40] Z. Zhang, J. Jia, B. Wang, and N. Z. Gong, "Backdoor attacks to graph neural networks," in *Proceedings of the 26th ACM symposium on access control models and technologies*, 2021, pp. 15–26.
- [41] J. Jia, Y. Liu, and N. Z. Gong, "Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 2043–2059.
- [42] J. Zhang, J. Chi, Z. Li, K. Cai, Y. Zhang, and Y. Tian, "Badmerging: Backdoor attacks against model merging," in *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, 2024, pp. 4450–4464.
- [43] Anthropic, "Claude 3.5 sonnet: Our first release in the claude 3.5 model family," <https://www.anthropic.com/news/claude-3-5-sonnet>, 2024, accessed: Nov. 9, 2025.
- [44] W. Zhou, X. Wang, L. Xiong, H. Xia, Y. Gu, M. Chai, F. Zhu, C. Huang, S. Dou, Z. Xi *et al.*, "Easyjailbreak: A unified framework for jailbreaking large language models," *arXiv preprint arXiv:2403.12171*, 2024.

- [45] Y. Yuan, W. Jiao, W. Wang, J.-t. Huang, P. He, S. Shi, and Z. Tu, "Gpt-4 is too smart to be safe: Stealthy chat with llms via cipher," *arXiv preprint arXiv:2308.06463*, 2023.
- [46] H. Lv, X. Wang, Y. Zhang, C. Huang, S. Dou, J. Ye, T. Gui, Q. Zhang, and X. Huang, "Codechameleon: Personalized encryption framework for jailbreaking large language models," *arXiv preprint arXiv:2402.16717*, 2024.
- [47] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv preprint arXiv:2307.15043*, 2023.
- [48] P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong, "Jailbreaking black box large language models in twenty queries," in *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 2025, pp. 23–42.
- [49] Y. Liu, Y. Jia, R. Geng, J. Jia, and N. Z. Gong, "Formalizing and benchmarking prompt injection attacks and defenses," in *33rd USENIX Security Symposium (USENIX Security 24)*, 2024, pp. 1831–1847.
- [50] F. Perez and I. Ribeiro, "Ignore previous prompt: Attack techniques for language models," *arXiv preprint arXiv:2211.09527*, 2022.
- [51] C. Sitawarin, N. Mu, D. Wagner, and A. Araujo, "Pal: Proxy-guided black-box attack on large language models," *arXiv preprint arXiv:2402.09674*, 2024.
- [52] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li *et al.*, "Harmbench: A standardized evaluation framework for automated red teaming and robust refusal," *arXiv preprint arXiv:2402.04249*, 2024.
- [53] P. Chao, E. Debenedetti, A. Robey, M. Andriushchenko, F. Croce, V. Schwag, E. Dobriban, N. Flammarion, G. J. Pappas, F. Tramèr *et al.*, "Jailbreakbench: An open robustness benchmark for jailbreaking large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 55 005–55 029, 2024.
- [54] Z. Zhu, B. Wu, Z. Zhang, and B. Wu, "Riskawarebench: Towards evaluating physical risk awareness for high-level planning of llm-based embodied agents," *arXiv e-prints*, pp. arXiv–2408, 2024.
- [55] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.
- [56] A. Liu, Z. Ying, L. Wang, J. Mu, J. Guo, J. Wang, Y. Ma, S. Liang, M. Zhang, X. Liu *et al.*, "Agentsafe: Benchmarking the safety of embodied agents on hazardous instructions," *arXiv preprint arXiv:2506.14697*, 2025.