Safety of Embodied Navigation: A Survey

Zixia Wang[1] , **Jia Hu**[1] , **Ronghui Mu**[1*]

[1]University of Exeter

{zw483,j.hu,r.mu2}@exeter.ac.uk

Abstract

As large language models (LLMs) continue to advance and gain influence, the development of embodied AI has accelerated, drawing significant attention, particularly in navigation scenarios. Embodied navigation requires an agent to perceive, interact with, and adapt to its environment while moving toward a specified target in unfamiliar settings. However, the integration of embodied navigation into critical applications raises substantial safety concerns. Given their deployment in dynamic, real-world environments, ensuring the safety of such systems is critical. This survey provides a comprehensive analysis of safety in embodied navigation from multiple perspectives, encompassing attack strategies, defense mechanisms, and evaluation methodologies. Beyond conducting a comprehensive examination of existing safety challenges, mitigation technologies, and various datasets and metrics that assess effectiveness and robustness, we explore unresolved issues and future research directions in embodied navigation safety. These include potential attack methods, mitigation strategies, more reliable evaluation techniques, and the implementation of verification frameworks. By addressing these critical gaps, this survey aims to provide valuable insights that can guide future research toward the development of safer and more reliable embodied navigation systems. Furthermore, the findings of this study have broader implications for enhancing societal safety and increasing industrial efficiency.

1 Introduction

In recent years, Large Language Models (LLMs) have garnered significant attention for their remarkable capabilities in perception, interaction, and reasoning. These advancements have contributed to the rise of Embodied Artificial Intelligence (Embodied AI), which serves as a bridge between the virtual and physical worlds. A key aspect of Embodied AI
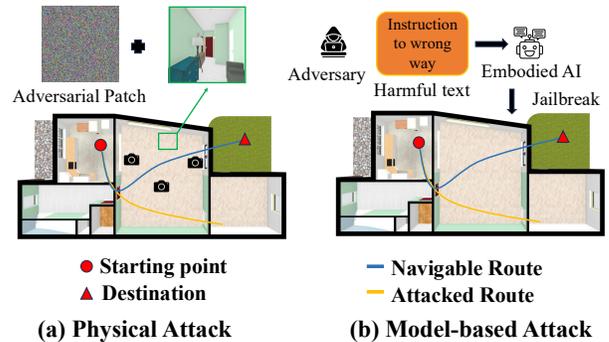
---

*Corresponding Author

Figure 1: Examples of two types of attacks. The red circle marks the starting point, the red triangle indicates the destination, the green line represents the navigable route, and the yellow line shows the attacked route. In the physical attack example, an adversarial patch disrupts navigation, while in the model-based attack example, a jailbreak injects harmful instructions, leading to incorrect actions.

is embodied navigation, which enables an AI agent to perceive and interact with its environment while moving toward a target or specified location in unfamiliar settings. This requires a combination of intelligent capabilities, including visual perception, mapping, planning, exploration, and reasoning. For example, consider an AI agent instructed to *"Retrieve a bottle of water from the kitchen fridge."* The agent must navigate to the kitchen, identify the fridge, pick up the correct item, and return to the designated location. Embodied navigation plays a crucial role in various real-world applications, including safety-critical scenarios such as robotic navigation [Wang *et al.*, 2024b] and autonomous driving [Li *et al.*, 2023]. Therefore, ensuring the safety and efficiency of embodied navigation is crucial.

However, the security of Embodied AI remains a significant concern, as it relies on deep neural networks (DNNs), which are susceptible to adversarial attacks [Liu *et al.*, 2020]. These vulnerabilities pose serious risks to the safety and reliability of embodied navigation systems. One type of attack alters the physical environment to mislead navigation perception. For example, adversarial patches or perturbations placed on objects or surfaces can mislead the input of the model, causing it to misunderstand its surroundings [Chen *et al.*, 2024]. Another form of attack directly targets the AI

model by injecting maliciously crafted inputs that manipulate its decision-making process [Liu *et al.*, 2024]. For example, carefully designed adversarial prompts can trick a large model into producing incorrect or harmful outputs. As a result, the agent may misidentify barriers, take incorrect paths, or even crash into objects (as shown in Figure 1, both physical patch attacks and model-based jailbreak attacks cause embodied navigation to deviate from its original correct path). While some existing studies focus on defense methods [Wu *et al.*, 2024a] or the construction of safety-related benchmarks [Yin *et al.*, 2024], a comprehensive survey on the safety of embodied navigation is still lacking, which limits a holistic understanding of the field's development.

In this paper, we aim to explore three key research questions: (1) *What risks do embodied navigation agents face?* (2) *What methods can be employed to mitigate these risks and enhance system reliability?* (3) *What metrics can be used to evaluate the safety of embodied navigation agents?* To address these questions, we present a comprehensive survey that summarizes recent advancements in three critical areas: **attacks, defenses, and evaluation** (as illustrated in Figure 2). First, after comprehensive investigation the safety risks, we categorize potential threats into two primary types: **physical attacks**, which are caused by environmental factors such as adversarial patches or lighting conditions, and **model-based attacks**, which exploit vulnerabilities in the navigation model itself, particularly in large-scale models. These attack types are further classified based on their nature, the attackers involved, and the methodologies employed. Next, we systematically examine existing defense mechanisms for embodied navigation, aligning them with the corresponding attack types to provide a structured understanding of adversarial threats and mitigation strategies. Additionally, we emphasize the critical role of well-structured datasets and robust evaluation metrics in ensuring the safety and reliability of embodied navigation systems. Given that evaluation is a fundamental aspect of safety assessment, we discuss the necessity of standardized benchmarks and comprehensive testing methodologies in advancing the security of embodied AI.

Building on our comprehensive review of existing research, we present some future directions in embodied AI safety. These include advancing attack strategies, particularly in multimodal AI settings; developing more effective and adaptive defense mechanisms for real-time navigation; and establishing standardized evaluation frameworks to enhance fairness and interpretability across diverse tasks. Moreover, we highlight the importance of verification techniques in quantifying robustness thresholds and defining theoretical performance bounds. We believe these efforts will contribute to the development of safer and more reliable embodied AI systems. Notably, [Zhang *et al.*, 2024c; Liu *et al.*, 2025] are works related to us, focusing primarily on embodied navigation (excluding safety aspects) and safety within embodied AI in one specific area (healthcare), respectively. In contrast, the central focus of our work is on safety issues within embodied navigation across a general domain. Our main contributions are as follows:

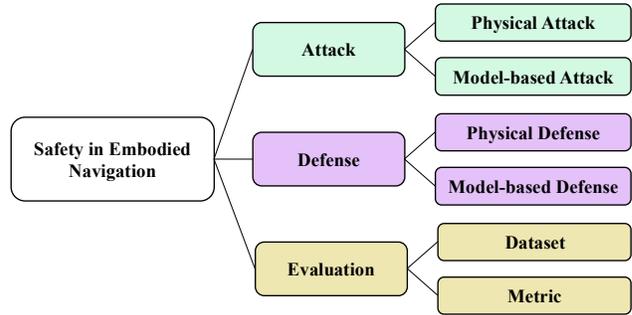- We provide a systematic review of attack and defense

Figure 2: Taxonomy: Safety of Embodied Navigation

methods related to the safety of embodied navigation. To the best of our knowledge, this is the first comprehensive study on the safety of embodied navigation.

- We compare and analyze recent evaluation datasets and metrics used for assessing embodied navigation safety.

- We present potential future research directions in embodied navigation safety to inform and inspire further advancements toward the development of more robust and reliable embodied navigation systems.

2 Preliminaries

In this section, we introduce the background knowledge in embodied navigation and safety to better illustrate the scope of this survey.

2.1 Overview of Embodied Navigation

Embodied AI refers to intelligent systems that integrate perception, decision-making, and action within a physical or simulated environment, enabling them to interact with and adapt to their surroundings autonomously. A crucial application of embodied AI is embodied navigation, which enables an agent to perceive its surroundings, plan a path, and move toward a target while adapting to dynamic and unfamiliar settings. In an embodied navigation task, an agent is placed in a visual environment E and given an instruction I to find a route R from a starting point S to a destination D. The route R is a sequence of viewpoints that the agent will follow. At each time step t, the agent observes multiple views $\{V_{t,i}\}$; some of these views indicate possible directions of movement. Using the instruction I, its previous observations, and previous commands $\{c_0, c_1, \ldots, c_{t-1}\}$, the agent selects the next command c_t. The process ends when the agent chooses the stop command, denoted as c_{stop}. Our survey builds upon and extends this foundational task.

Embodied navigation focuses on enabling agents to move and interact in physical environments using visual and sensory input. Key tasks include object goal navigation (e.g., locating specific objects like "kitchen" in unknown settings [Chen *et al.*, 2024; Ying *et al.*, 2023; Yang *et al.*, 2024]), image goal navigation (reaching a target depicted in an image like "fridge" [Mezghan *et al.*, 2022]) , visual language navigation (VLN) (following natural language instructions to

traverse a route [Jiao *et al.*, 2024; Lu *et al.*, 2024]) and interactive navigation (answering questions about the environment [Liu *et al.*, 2024]).

2.2 Safety in Embodied Intelligence

Currently, research on embodied AI safety remains relatively limited, with different studies using various metrics to assess these systems. In general, safety is understood as a model's ability to withstand perturbations, where greater robustness means stronger resistance to disruption. Specifically, adversarial perturbations pose a significant challenge to embodied navigation by distorting perception and decision-making. At time step t, the perturbed observation is defined as $V'_{t,i} = V_{t,i} + \delta_{t,i}$, where $\delta_{t,i}$ represents an adversarial perturbation applied to the original observation $V_{t,i}$, which can alter the agent's perception of the environment. Humans are usually unable to detect these disturbances. Given an input $V_{t,i}$, a model F typically produces an output c_t that aligns with human expectations. However, when $V'_{t,i}$ is presented instead, the model's response may deviate significantly, leading to unexpected or even harmful consequences. As a result, instead of selecting the correct command c_t based on the original observations, the agent chooses c'_t under the perturbed observations, where $c'_t \neq c_t$. This incorrect decision may cause the agent to deviate from its intended route, leading to navigation failure and potentially unsafe outcomes.

Embodied intelligence systems face two major security threats: physical threats and model threats. Since deep neural networks (DNNs) are highly sensitive to such disruptions, external factors such as malicious patches [Ying *et al.*, 2023; Chen *et al.*, 2024] or changes in lighting conditions [Zhang *et al.*, 2024b; Sun *et al.*, 2024] can distort the system's understanding of its surroundings, leading to navigation errors. On the other hand, model threats arise from vulnerabilities within large language models (LLMs). Attackers may inject malicious instructions or exploit hallucination effects [Jiao *et al.*, 2024; Wang *et al.*, 2024a; Zhang *et al.*, 2024a; Huang *et al.*, 2024a; Dong *et al.*, 2024], potentially causing the system to generate inaccurate or even risky route plans. A detailed discussion of these threats will be presented in Chapter 3.

3 Attack

In this section, we examine attacks on embodied navigation, broadly classified into two categories: physical attacks and model-based attacks, with some attacks combining elements of both. Table 1 presents different types of representative attacks. However, research on the vulnerabilities of embodied systems remains limited. To address this gap, we discuss potential future attack vectors in Chapter 6.

3.1 Physical Attack

Since the introduction of achievable physical adversarial samples [Kurakin *et al.*, 2018], physical adversarial attacks widely applied in various computer vision tasks, such as facial recognition [Wei *et al.*, 2024] and traffic sign detection [Suryanto *et al.*, 2023].

However, research on adversarial attacks targeting embodied navigation agents remains relatively limited. To establish a foundation for understanding these threats, we first review physical patch attacks in 2D pixel spaces before exploring their extensions to 3D settings. The early exploration of physical attacks was initiated by [Brown *et al.*, 2017], who proposed a method for generating universal adversarial image patches capable of attacking any scene. Later, [Chen *et al.*, 2018] introduced adversarial perturbations on stop signs, causing Faster R-CNN to misclassify them and threatening autonomous vehicles. Furthermore, some studies [Wei *et al.*, 2023; Li and Ji, 2021] focused on optimizing patch placement to enhance attack effectiveness. In embodied navigation attacks, the attack is achieved by [Ying *et al.*, 2023] through the application of agnostic perturbations to each input frame.

Due to the inherent limitations of 2D patches, researchers have redirected their attention to investigating the impact of patch-based attacks on 3D properties, such as rotation and translation [Zeng *et al.*, 2019]. [Liu *et al.*, 2020] was an early work on attacks against embodied navigation agents, which focused on the physical attributes of objects in key scene views, such as textures and 3D shapes. Meanwhile, [Athalye *et al.*, 2018; Wiyatno and Xu, 2019] introduced the Expectation Over Transformation (EOT) method, which generated robust 3D adversarial examples by simulating real-world variations, including rotation, scaling, and blurring.

Influenced by methods such as EOT, numerous attacks have emerged. [Xu *et al.*, 2020] proposed attaching adversarial patches to t-shirts to generate robust adversarial samples that simulate deformation effects. Similarly, [Yang *et al.*, 2024] explored the appearance optimization of 3D adversarial objects, aiming to manipulate how objects are perceived in the environment and induce the desired behavior in a pretrained VLN agent, thereby achieving an attack in embodied navigation. Additionally, some methods explored 3D adversarial camouflages by disrupting object textures within navigation scenes, particularly in autonomous driving. [Huang *et al.*, 2024b] developed more effective and transferable techniques for generating targeted 3D adversarial examples. [Suryanto *et al.*, 2023] proposed camouflage attack methods based on texture rendering, while [Wang *et al.*, 2021] implemented attacks on the full 3D surface of vehicles.

Early research on attacks against embodied navigation agents was constrained by viewpoint variations and environmental complexity, limiting their effectiveness in real-world settings. To address these challenges, recent studies have developed specialized attack strategies tailored for navigation tasks. Adversarial textures for clothing [Hu *et al.*, 2023b] were initially explored to enhance robustness against viewpoint variations. Similarly, [Chen *et al.*, 2024] also leveraged a multi-view approach, proposing a method that attaches adversarial patches with learnable textures and opacities to objects, integrating multiple viewpoints to achieve physical attacks.

In addition to patch-based physical attacks, some adversarial attack strategies targeting navigation shifted their focus to adversarial light. Initially, this approach was applied to image classification, where a projector altered physical light condi-

Type	Paper	Attacker	Attack Type	Technique	Brief Description
Physical Attack	[Ying *et al.*, 2023]	User	White-box	2D-patch	Constant image-agnostic perturbation applied to each input frame
	[Wei *et al.*, 2023]	User	Black-box	2D-patch	Optimizing adversarial patch placement for optimal positioning
	[Suryanto *et al.*, 2023]	User	Black-box	3D-patch	General-purpose adversarial patches for vehicle applications
	[Yang *et al.*, 2024]	User	White-box	3D-patch	Optimized the appearance of 3D adversarial objects
	[Huang *et al.*, 2024b]	User	Black-box	3D-patch	Transferable targeted 3D adversarial examples
	[Hu *et al.*, 2023b]	User	White-box	Multi-view-patch	Adversarial texture for clothes
	[Chen *et al.*, 2024]	User	White-box	Multi-view-patch	Multiview optimization strategy based on object-aware sampling
	[Hu *et al.*, 2023a]	User	Black-box	Adversarial Light	Optimizes the physical parameters of laser spots to perform physical attacks
	[Zhang *et al.*, 2024b]	Third Party	White-box	Adversarial Light	Electromagnetic Signal Injection Attacks
	[Sun *et al.*, 2024]	User, Third Party	Black-box	Adversarial Light	Dynamically tailored non-contact laser attack
Model-based Attack	[Sun *et al.*, 2024]	User, Third Party	Black-box	Reinforcement Learning	Dynamically tailored non-contact laser attack
	[Zhang *et al.*, 2022]	User	Black-box	Federated Learning	Backdoor attack on FL-based embodied agents
	[Jiao *et al.*, 2024]	User	White-box	LLM-Backdoor attack	Backdoor attacks against embodied LLM-based decision-making systems
	[Wang *et al.*, 2024a]	User, Third Party	Black-box	LLM-Backdoor attack	Backdoor attacks on VLM-based robotic manipulation
	[Zhang *et al.*, 2024a]	Third Party	Black-box	LLM-Jailbreak attack	Manipulation, alignment, and knowledge for attacks
	[Lu *et al.*, 2024]	User	Black-box	LLM-Jailbreak attack	Adversarial and meaningful suffixes with a focus on simple words
	[Liu *et al.*, 2024]	User	White-box	LLM-Jailbreak attack	Targeted attacks for controlled manipulation and untargeted attacks for random disruptions

Table 1: Different types of representative attacks: "Attacker" refers to the adversary, categorized into user and third party; "Attack type" is classified into black-box and white-box attacks, while "Technique" corresponds to different attack methodologies.

tions to deceive classifiers [Huang and Ling, 2022]. Later, adversarial light techniques were extended to embodied navigation. Geometric light attacks distorted entire images, leading to the misinterpretation of navigation signs by vehicles [Hu *et al.*, 2023a]. Similarly, electromagnetic signal injection manipulated visual inputs, affecting both classification and navigation tasks [Zhang *et al.*, 2024b]. Furthermore, laser emitters were used to attack embodied navigation agents by exploiting vulnerabilities in their perception systems [Sun *et al.*, 2024].

3.2 Model-based Attack

In model-based attacks, unlike physical attacks that focus on changes in the environment, these attacks are directed specifically at the model itself.

In embodied navigation tasks, reinforcement learning (RL) can be employed to train agents on how to navigate effectively. Similar to models trained in standard gaming environments, embodied navigation systems are also susceptible to certain attacks based on reinforcement learning [Mu *et al.*, 2024]. Rather than using time-consuming heuristic algorithms, [Sun *et al.*, 2024] employed reinforcement learning to optimize adversarial laser attack strategies, improving efficiency.

Federated Learning (FL) enables multiple clients, such as household navigation environments, to collaboratively train navigation models without sharing raw data with a central server, thereby preserving data privacy. In the context of embodied navigation, FL facilitates decentralized learning, allowing agents to adapt to different environments while maintaining data security. However, the decentralized nature of FL also introduces security risks. The opacity of local training processes makes the system vulnerable to adversarial manipulation [Lyu *et al.*, 2022]. To explore these vulnerabilities, [Zhang *et al.*, 2022] investigated how malicious clients could manipulate their local training data, allowing attackers to control the global model under specific conditions.

Large Language Models (LLMs) demonstrated immense potential for navigation in embodied artificial intelligence. With extensive common sense and advanced reasoning capabilities, these models enabled robots to better comprehend complex language commands and execute high-level tasks

with greater understanding and adaptability [Sharan *et al.*, 2023]. However, large models also faced several security issues, primarily jailbreak attacks and backdoor attacks. Jailbreak attacks exploit model vulnerabilities to bypass safety mechanisms, allowing attackers to generate restricted content using crafted prompts. Backdoor attacks embed hidden triggers, making the model behave maliciously when specific inputs are given. Research on backdoor attacks explored various mechanisms to compromise LLMs. Hidden triggers were implanted into models using word-based, scene-based, and Retrieval-Augmented Generation (RAG) techniques, demonstrating how these methods could embed vulnerabilities into systems [Jiao *et al.*, 2024]. Additionally, visual-language models (VLMs) integrated into robotic systems were examined for potential exploits, revealing that adversaries could manipulate them to execute harmful actions in real-world environments [Wang *et al.*, 2024a]. In addition, jailbreak attacks in embodied LLM-based robots were achieved through voice-based user interactions, effectively bypassing safety and ethical constraints [Zhang *et al.*, 2024a]. Another approach involved generating adversarial and meaningful simple word suffixes to influence embodied AI, enabling precise voice injections capable of causing harm in the physical world, affecting both environments and humans [Lu *et al.*, 2024]. Also, both untargeted and targeted attacks were employed to execute jailbreaks in LLM-based embodied models, further highlighting their security vulnerabilities [Liu *et al.*, 2024].

4 Defense

In this section, we explore various defense mechanisms for embodied navigation, which are generally classified into two main categories: physical defenses and model-based defenses. In particular, some defenses integrate elements from both categories. Table 2 provides an overview of representative defense types.

4.1 Physical Defense

Several prior studies explored defenses against patch attacks on pixels instead of specifically designing for embodied systems, employing both empirical strategies [Xu *et al.*, 2023; Wu *et al.*, 2024b] and certified approaches [Xiang *et al.*,

2021; Xiang *et al.*, 2024]. One approach focused on "detection and removal", where adversarial purification was applied to mitigate the impact of adversarial patches [Xu *et al.*, 2023]. Another method introduced a detection framework targeting naturalistic adversarial patches with deceptive features [Wu *et al.*, 2024b]. Beyond empirical strategies, a small receptive field CNN was used to limit the number of features that adversarial patches could corrupt, thereby improving model robustness [Xiang *et al.*, 2021]. Subsequent work further enhanced both efficacy and robustness by refining these certified defenses [Xiang *et al.*, 2024].

Unlike previous passive defenses, active defense mechanisms were introduced for embodied navigation, utilizing recurrent feedback to actively counter adversarial patches [Wu *et al.*, 2024a]. This approach leveraged environmental context, addressing misaligned adversarial patches in real-world 3D settings.

4.2 Model-based Defense

We name "model-based defense" to align with model-based attacks, as the defenses discussed here are specifically designed to counter the previously mentioned attacks. By maintaining this alignment, these defense strategies leverage model-driven mechanisms to effectively mitigate adversarial threats. The Embodied Active Defense (EAD) method was introduced to tackle adversarial patches in the 3D real world, actively integrating perception and action to interact with and adapt to the environment, thereby enhancing decision-making [Wu *et al.*, 2024a]. To assess the safety of federated embodied agents, a real-time defense mechanism was developed by [Zhang *et al.*, 2022], implementing a Prompt-Based Aggregation (PBA) mechanism that detects malicious clients by analyzing vision-language alignment variance, thus providing more robust protection against federated learning attacks. Some studies focused on defense strategies for embodied navigation based on large language models. Various methods were evaluated to determine their effectiveness against backdoor attacks on embodied models [Jiao *et al.*, 2024]. Notably, directly deploying defense models (such as Llama-Guard-2, Llama-Guard-3, and Harmbench) has been a common approach. Both prompt-level and model-level defenses were explored to mitigate jailbreak attacks on embodied AI [Lu *et al.*, 2024].

5 Evaluation

In this section, we focus on the safety assessment issues in embodied navigation. Initially, we review safety-related datasets, followed by an organization of metrics used to evaluate safety.

5.1 Dataset

Some of the benchmarks used in our work in Chapters 3 and 4 were not originally designed for embodied navigation (e.g. [Chen *et al.*, 2024]). Additionally, some works have created their own datasets to meet their specific experimental needs [Yang *et al.*, 2024]. In this section, we concentrate on the recent and representative datasets for the safety of embodied navigation. And we categorize the benchmarks based on the

Defense	Physical Defense	Model-based Defense	Attack Type	Core Method
[Xu *et al.*, 2023]	✓		white-box	Detection and Remove
[Wu *et al.*, 2024b]	✓		black-box	Feature Aligned Learning
[Wu *et al.*, 2024a]	✓	✓	white-box	Reinforcement learning
[Zhang *et al.*, 2022]		✓	black-box	Federated learning
[Lu *et al.*, 2024]		✓	black-box	LLM Jailbreak
[Jiao *et al.*, 2024]		✓	white-box	LLM Backdoor Attack

Table 2: Different types of representative defenses. "Attack Type" refers to the category of attacks that this defense is designed to counter.

number of model parameters into those designed for classic models and those designed for LLMs.

Datasets for classic models
Based on the work of [Li *et al.*, 2023], a physical attack naturalness dataset was constructed using human ratings and gaze data. Due to the limitations of virtual environments in replicating object interactivity and scene scale found in real-world settings, a photo-based 3D benchmark was later developed [Kim *et al.*, 2024]. By integrating authentic scenes, objects, and room layouts, this benchmark allowed agents to better comprehend language instructions, complete household tasks, and operate in large-scale, multi-room real-world environments. Efforts to enhance embodied navigation benchmarks extended to diverse settings, including houses, gardens, restaurants, and offices. Objects within these environments were annotated with detailed physical and semantic attributes, with a focus on both reinforcement learning (RL) agents and safety concerns [Li *et al.*, 2024]. Further advancements in multimodal lifelong navigation introduced a benchmark designed to challenge agents in open-vocabulary navigation tasks [Khanna *et al.*, 2024]. Agents were required to locate targets specified by category names, natural language descriptions, or images, contributing to the development of general-purpose navigation systems .

Datasets for LLMs
With the development of LLMs, embodied LLM agents became more effective in interacting with people and making informed decisions in navigation. However, as previously discussed, LLMs remain vulnerable to jailbreak and backdoor attacks. In response, various benchmarks were introduced to evaluate LLM-based navigation systems, particularly focusing on recently developed datasets. A benchmark suite was designed to automatically assess the task-planning capabilities of LLMs [Choi *et al.*, 2024]. Each dataset sample provided natural language instructions and an environment to the planner. The simulator executed the planned actions and evaluated performance by comparing the final state with a predefined target condition. While its contributions, the dataset primarily focused on planning abilities rather than safety concerns.

To explore physical risks in embodied AI, dangerous scenarios were generated using LLMs and diffusion models, leading to an automated framework for risk assessment [Zhu *et al.*, 2024]. Various open-source and closed-source models were evaluated within this framework. However, safety analysis was mostly restricted to the input text, treating the embodied environment as an additional input rather than a core

Evaluation Dataset	Dataset Construction	# Size	Evaluation Metric	Domain
[Li *et al.*, 2023]	Autonomous driving image	2,688 images	Human-based	Physical world attacks
[Kim *et al.*, 2024]	ALFRED[1]&Human annotation	150 scenes	Formula-based	Photo-realistic environments navigation
[Khanna *et al.*, 2024]	Real-world 3D scans from HM3DSem[2]	312 categories	Formula-based	Multi-modal lifelong navigation
[Choi *et al.*, 2024]	ALFRED&WAH[3]	308 tasks	Formula-based	Language oriented task planner
[Zhu *et al.*, 2024]	GPT-4o generation	2,636 samples	Formula-based	Physical risk task planning
[Yin *et al.*, 2024]	GPT-4 generation	750 tasks	Model-based	Safe task planning
[Wang *et al.*, 2024b]	GPT-4 generation&Holodeck	4,614 scenes	Formula-based	LVLMS for object navigation
[Wang *et al.*, 2024a]	GPT-4 generation&Human annotation	328 tasks	Formula-based	MLLM navigation

Table 3: Different types of representative evaluation datasets can be described by several aspects: "Dataset Construction" details the process used to build the dataset; "# Size" indicates the size of the dataset; "Evaluation Metric" represents the three different classification types; and "Domain" denotes the scope within which the dataset is applied.

aspect of evaluation. Recognizing this limitation, an alternative approach placed safety concerns at the center of evaluation, focusing on embodied agents that directly interact with the physical world rather than language models that only process text. [Yin *et al.*, 2024] addressed ten common risks affecting humans and property, categorizing tasks into detailed tasks, abstract tasks, and long-horizon tasks to explore safety issues at various levels of abstraction and task duration.

Efforts to expand navigation-related datasets included the introduction of tasks requiring agents to navigate to various target objects across multiple scenarios. A dataset [Wang *et al.*, 2024b] covering 4,614 houses across 81 scenario types was constructed, utilizing Holodeck[4] to generate textual descriptions of houses, while GPT-4 was employed to determine layout, style, and object placement. Textual annotations were later added to enrich the dataset's contextual information. In the realm of multimodal large-model embodied agents, an evaluation framework was established to assess capabilities across five different categories, including navigation [Cheng *et al.*, 2025]. Task generation was powered by LLMs, with manual annotation and rigorous scene screening ensuring the dataset's high quality and reliability. Notably, unlike traditional dataset creation methods, leveraging the generative power of LLMs in combination with human screening has emerged as a promising direction for developing high-quality embodied AI datasets.

5.2 Metric

The evaluation approaches for embodied navigation are diverse. Depending on the dataset and specific tasks, the methods can vary. Here, we primarily categorize the evaluation methods into three groups: human-based evaluation, formula-based evaluation, and model-based evaluation.

Human-based evaluation

Human-based evaluation is an assessment method that directly involves human judgment in evaluating a system's performance. It is the simplest evaluation approach, ensuring both accuracy and reliability. [Li *et al.*, 2023] investigated the naturalness of physical-world attacks using human ratings and gaze data. Manual annotation was widely

used in most studies to ensure the high quality of datasets, as demonstrated by [Kim *et al.*, 2024; Cheng *et al.*, 2025; Yin *et al.*, 2024]. In some works, human evaluation was directly used to determine the correctness of the results (e.g., [Huang *et al.*, 2022]), with the *success rate* serving as the primary evaluation metric.

Formula-based evaluation

Due to the costly and time-consuming nature of human-based evaluation, most benchmarks have begun shifting towards formula-based evaluation methods. This approach relies on predefined formulas and definitions to conduct the assessment. Here, we introduce some commonly used methods and formulas. In the benchmarks we mentioned, common metrics include: Success Rate (SR) [Yin *et al.*, 2024; Kim *et al.*, 2024], Success weighted by Path Length (SPL) [Wang *et al.*, 2024b; Khanna *et al.*, 2024], Success weighted by Episode Length (SEL) [Wang *et al.*, 2024b], and Goal-condition Success (GcS, abbreviated as GC) [Cheng *et al.*, 2025; Kim *et al.*, 2024].

An episode is considered successful if the target object appears in the agent's egocentric view and is within 1.5 meters of the agent. To maintain consistent notation, we denote the total number of episodes by M and index each episode by k (where $k = 1, 2, \ldots, M$). In this framework, s_k is a binary indicator of success (with $s_k = 1$ if the episode is successful, and $s_k = 0$ otherwise), d_k represents the length of the optimal (i.e., shortest) path to the target, and p_k is the length of the path traversed by the agent. For metrics based on episode length, d_k^a and p_k^a denote the number of actions along the optimal path and the agent's actual trajectory, respectively, while c_k indicates the number of goal conditions satisfied in episode k, and C is the total number of predefined goal conditions. Using these definitions, the metrics are computed as follows: the SR is given by $SR = \frac{1}{M}\sum_{k=1}^{M} s_k$; the SPL is calculated as $SPL = \frac{1}{M}\sum_{k=1}^{M} s_k \cdot \frac{d_k}{\max(d_k, p_k)}$; the SEL is determined by $SEL = \frac{1}{M}\sum_{k=1}^{M} s_k \cdot \frac{d_k^a}{\max(d_k^a, p_k^a)}$; and the GC is computed as $GC = \frac{1}{M}\sum_{k=1}^{M} \frac{c_k}{C}$. If a system is considered safe, the SR, SPL, SEL, and GC metrics should be as high as possible, reflecting its ability to perform tasks efficiently, securely, and reliably. When considering navigation efficiency simultaneously, time t is also an important evaluation metric.

---

[1] https://github.com/askforalfred/alfred

[2] https://aihabitat.org/datasets/hm3d-semantics/

[3] https://github.com/xavierpuigf/watch_and_help

[4] https://yueyang1996.github.io/holodeck/

Model-based evaluation

Beyond leveraging large models for data generation, some benchmarks also adopt model-based evaluation for assessing performance. In particular, abstract tasks often allow for multiple valid execution strategies rather than a single definitive solution. To address this variability, [Yin *et al.*, 2024] employed GPT-4 to evaluate the plausibility and effectiveness of execution plans generated by the model. This approach ensures that the proposed plans align with task objectives and maintain coherence. Building on these evaluations, performance was further quantified by computing both the success rate and the probability of rejection, providing a systematic way to assess the model's robustness and safety.

6 Future Directions

In this section, we discuss several unresolved challenges in the safety of embodied navigation, highlighting key issues that remain to be addressed. We aim to offer valuable insights and propose potential research directions that could foster the development of safer and more efficient embodied navigation systems in the future.

6.1 Potential Attack Methods

There are three potential research directions for adversarial attacks: **(a) Enhancing Robustness in Dynamic Environments**: Some existing studies primarily focus on developing attack and defense strategies for specific agent tasks, demonstrating the effectiveness of their proof-of-concept approaches. However, these methods often face significant limitations when applied to real-world, complex environments. For instance, object-trigger-based attacks require precise visual consistency across multiple viewpoints, making them less effective in dynamic settings with varying perspectives. **(b) Expanding Attack Types**: Certain attack strategies operate under a black-box assumption, where the attacker lacks prior knowledge of the model's internal mechanisms, thereby limiting their applicability in scenarios that require more fine-grained adversarial optimization. As a result, expanding the scope of attack methodologies, including white-box attacks, remains a critical avenue for further research. **(c) Adversarial Attacks on Multimodal Models**: Existing model-based attack research has primarily focused on large language models. However, with the rapid advancement of multimodal large models, traditional attack paradigms may not seamlessly transfer to multimodal settings. Therefore, investigating attack strategies specifically designed for multimodal models, such as cross-modal perturbations, represents an important yet unsolved direction.

6.2 Robust Defense Strategies

Currently, research on physical defenses (e.g., patch-based defenses) in embodied navigation remains limited, leaving significant room for further exploration. Existing defense mechanisms are often designed for other tasks and may not fully address the unique challenges of embodied navigation, such as real-time decision-making and continuous interaction with dynamic environments. In LLM-based navigation systems, models may inherit vulnerabilities from text-based models, including sensitivity to prompts and susceptibility to adversarial text. Strengthening the defense capabilities of LLMs in embodied navigation, such as developing more robust language understanding mechanisms, integrating adversarial-resistant knowledge injection, or enforcing multimodal consistency constraints, remains a critical research direction. Additionally, due to the real-time interactive nature of embodied systems, runtime monitoring techniques could be explored as a defense strategy to dynamically identify and counteract adversarial threats during execution.

6.3 Reliable Evaluation

Current research often focuses primarily on qualitative assessments without conducting rigorous quantitative experiments. Even studies that incorporate quantitative evaluations typically rely on different datasets or purpose-built datasets for experimentation, lacking direct comparisons with other security-related approaches. As a result, a more systematic analysis of security remains an important research direction. Moreover, evaluation methods can be transitioned to model-based approaches, such as leveraging GPT-4 or other large models to assess accuracy, which can automate the process and reduce human effort. Additionally, different AI tasks, such as visual exploration and LLM-based question answering, adopt distinct evaluation metrics, making direct comparisons between studies challenging. Therefore, future research should focus on developing a more unified evaluation framework to ensure fairness and interpretability across different tasks. Furthermore, exploring new evaluation methods, such as integrating multiple metrics or incorporating human feedback, could further enhance the reliability and quality of AI task evaluation.

6.4 Verification Techniques

Current research lacks a systematic approach to the verification of embodied navigation, making this a promising direction for further exploration. One potential avenue is quantifying the range of input perturbations that do not affect the model's output, thereby establishing robustness thresholds. Additionally, computing theoretical bounds under different conditions presents another valuable research direction, as these bounds not only provide guidance for safety research but also serve as key metrics for evaluating system robustness. These efforts would contribute to enhancing the reliability and security of embodied navigation while providing theoretical support for the development of more robust navigation algorithms.

7 Conclusion

In this paper, we present a detailed overview of the safety of embodied navigation. We review recent research advancements from three key perspectives: attack strategies, defense mechanisms, and evaluation methodologies. Finally, based on the current state of research, we identify several promising directions for future investigation, aiming to foster the development of safer and more robust embodied navigation systems.

References

[Athalye *et al.*, 2018] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *ICML*, 2018.

[Brown *et al.*, 2017] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv*, 2017.

[Chen *et al.*, 2018] Shang-Tse Chen, Cory Cornelius, Jason Martin, and Duen Horng (Polo) Chau. Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector. In *ECML PKDD*, 2018.

[Chen *et al.*, 2024] Meng Chen, Jiawei Tu, Chao Qi, Yonghao Dang, Feng Zhou, Wei Wei, and Jianqin Yin. Towards physically-realizable adversarial attacks in embodied vision navigation. *arXiv preprint arXiv:2409.10071*, 2024.

[Cheng *et al.*, 2025] Zhili Cheng, Yuge Tu, Ran Li, Shiqi Dai, Jinyi Hu, Shengding Hu, Jiahao Li, Yang Shi, Tianyu Yu, Weize Chen, et al. Embodiedeval: Evaluate multi-modal llms as embodied agents. *arXiv*, 2025.

[Choi *et al.*, 2024] Jae-Woo Choi, Youngwoo Yoon, Hyobin Ong, Jaehong Kim, and Minsu Jang. Lota-bench: Benchmarking language-oriented task planners for embodied agents. *arXiv*, 2024.

[Dong *et al.*, 2024] Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, et al. Safeguarding large language models: A survey. *arXiv*, 2024.

[Hu *et al.*, 2023a] Chengyin Hu, Yilong Wang, Kalibinuer Tiliwalidi, and Wen Li. Adversarial laser spot: Robust and covert physical-world attack to dnns. In *ACML*, 2023.

[Hu *et al.*, 2023b] Zhanhao Hu, Wenda Chu, Xiaopei Zhu, Hui Zhang, Bo Zhang, and Xiaolin Hu. Physically realizable natural-looking clothing textures evade person detectors via 3d modeling. In *CVPR*, 2023.

[Huang and Ling, 2022] Bingyao Huang and Haibin Ling. Spaa: Stealthy projector-based adversarial attacks on deep image classifiers. In *IEEE VR*, 2022.

[Huang *et al.*, 2022] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, et al. Inner monologue: Embodied reasoning through planning with language models. *arXiv*, 2022.

[Huang *et al.*, 2024a] Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, 2024.

[Huang *et al.*, 2024b] Yao Huang, Yinpeng Dong, Shouwei Ruan, Xiao Yang, Hang Su, and Xingxing Wei. Towards transferable targeted 3d adversarial attack in the physical world, 2024.

[Jiao *et al.*, 2024] Ruochen Jiao, Shaoyuan Xie, Justin Yue, Takami Sato, Lixu Wang, Yixuan Wang, Qi Alfred Chen, and Qi Zhu. Exploring backdoor attacks against large language model-based decision making. *arXiv*, 2024.

[Khanna *et al.*, 2024] Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Theophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. Goat-bench: A benchmark for multi-modal lifelong navigation. In *CVPR*, 2024.

[Kim *et al.*, 2024] Taewoong Kim, Cheolhong Min, Byeonghwi Kim, Jinyeon Kim, Wonje Jeung, and Jonghyun Choi. Realfred: An embodied instruction following benchmark in photo-realistic environments. In *ECCV*, 2024.

[Kurakin *et al.*, 2018] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[Li and Ji, 2021] Xiang Li and Shihao Ji. Generative dynamic patch attack, 2021.

[Li *et al.*, 2023] Simin Li, Shuning Zhang, Gujun Chen, Dong Wang, Pu Feng, Jiakai Wang, Aishan Liu, Xin Yi, and Xianglong Liu. Towards benchmarking and assessing visual naturalness of physical world adversarial attacks. In *CVPR*, 2023.

[Li *et al.*, 2024] Chengshu Li, Ruohan Zhang, Josiah Wong, Cem Gokmen, Sanjana Srivastava, Roberto Martín-Martín, Chen Wang, Gabrael Levine, Wensi Ai, Benjamin Martinez, et al. Behavior-1k: A human-centered, embodied ai benchmark with 1,000 everyday activities and realistic simulation. *arXiv*, 2024.

[Liu *et al.*, 2020] Aishan Liu, Tairan Huang, Xianglong Liu, Yitao Xu, Yuqing Ma, Xinyun Chen, Stephen J. Maybank, and Dacheng Tao. Spatiotemporal attacks for embodied agents, 2020.

[Liu *et al.*, 2024] Shuyuan Liu, Jiawei Chen, Shouwei Ruan, Hang Su, and Zhaoxia Yin. Exploring the robustness of decision-level through adversarial attacks on llm-based embodied models. In *ICM*, 2024.

[Liu *et al.*, 2025] Yihao Liu, Xu Cao, Tingting Chen, Yankai Jiang, Junjie You, Minghua Wu, Xiaosong Wang, Mengling Feng, Yaochu Jin, and Jintai Chen. From screens to scenes: A survey of embodied ai in healthcare. *arXiv*, 2025.

[Lu *et al.*, 2024] Xuancun Lu, Zhengxian Huang, Xinfeng Li, Wenyuan Xu, et al. Poex: Policy executable embodied ai jailbreak attacks. *arXiv*, 2024.

[Lyu *et al.*, 2022] Lingjuan Lyu, Han Yu, Xingjun Ma, Chen Chen, Lichao Sun, Jun Zhao, Qiang Yang, and S Yu Philip. Privacy and robustness in federated learning: Attacks and defenses. *TNNLS*, 2022.

[Mezghan *et al.*, 2022] Lina Mezghan, Sainbayar Sukhbaatar, Thibaut Lavril, Oleksandr Maksymets, Dhruv Batra, Piotr Bojanowski, and Karteek Alahari. Memory-augmented reinforcement learning for image-goal navigation. In *IROS*, 2022.

[Mu *et al.*, 2024] Ronghui Mu, Leandro Soriano Marcolino, Yanghao Zhang, Tianle Zhang, Xiaowei Huang, and Wenjie Ruan. Reward certification for policy smoothed reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21429–21437, 2024.

[Sharan *et al.*, 2023] SP Sharan, Francesco Pittaluga, Manmohan Chandraker, et al. Llm-assist: Enhancing closed-loop planning with language-based reasoning. *arXiv*, 2023.

[Sun *et al.*, 2024] Yitong Sun, Yao Huang, and Xingxing Wei. Embodied laser attack: Leveraging scene priors to achieve agent-based robust non-contact attacks. In *ICM*, 2024.

[Suryanto *et al.*, 2023] Naufal Suryanto, Yongsu Kim, Harashta Tatimma Larasati, Hyoeun Kang, Thi-Thu-Huong Le, Yoonyoung Hong, Hunmin Yang, Se-Yoon Oh, and Howon Kim. Active: Towards highly transferable 3d physical camouflage for universal and robust vehicle evasion, 2023.

[Wang *et al.*, 2021] Donghua Wang, Tingsong Jiang, Jialiang Sun, Weien Zhou, Xiaoya Zhang, Zhiqiang Gong, Wen Yao, and Xiaoqian Chen. Fca: Learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack, 2021.

[Wang *et al.*, 2024a] Xianlong Wang, Hewen Pan, Hangtao Zhang, Minghui Li, Shengshan Hu, Ziqi Zhou, Lulu Xue, Peijin Guo, Yichen Wang, Wei Wan, et al. Trojan-robot: Physical-world backdoor attacks against vlm-based robotic manipulation. *arXiv*, 2024.

[Wang *et al.*, 2024b] Zhaowei Wang, Hongming Zhang, Tianqing Fang, Ye Tian, Yue Yang, Kaixin Ma, Xiaoman Pan, Yangqiu Song, and Dong Yu. Divscene: Benchmarking lvlms for object navigation with diverse scenes and objects. *arXiv preprint arXiv:2410.02730*, 2024.

[Wei *et al.*, 2023] Xingxing Wei, Shouwei Ruan, Yinpeng Dong, and Hang Su. Distributional modeling for location-aware adversarial patches, 2023.

[Wei *et al.*, 2024] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *NeurIPSs*, 2024.

[Wiyatno and Xu, 2019] Rey Reza Wiyatno and Anqi Xu. Physical adversarial textures that fool visual object tracking, 2019.

[Wu *et al.*, 2024a] Lingxuan Wu, Xiao Yang, Yinpeng Dong, Liuwei Xie, Hang Su, and Jun Zhu. Embodied active defense: Leveraging recurrent feedback to counter adversarial patches. *arXiv*, 2024.

[Wu *et al.*, 2024b] Siyang Wu, Jiakai Wang, Jiejie Zhao, Yazhe Wang, and Xianglong Liu. Napguard: Towards detecting naturalistic adversarial patches. In *CVPR*, 2024.

[Xiang *et al.*, 2021] Chong Xiang, Arjun Nitin Bhagoji, Vikash Sehwag, and Prateek Mittal. {PatchGuard}: A provably robust defense against adversarial patches via small receptive fields and masking. In *USENIX*, 2021.

[Xiang *et al.*, 2024] Chong Xiang, Tong Wu, Sihui Dai, Jonathan Petit, Suman Jana, and Prateek Mittal. {PatchCURE}: Improving certifiable robustness, model utility, and computation efficiency of adversarial patch defenses. In *USENIX*, 2024.

[Xu *et al.*, 2020] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world, 2020.

[Xu *et al.*, 2023] Ke Xu, Yao Xiao, Zhaoheng Zheng, Kaijie Cai, and Ram Nevatia. Patchzero: Defending against adversarial patch attacks by detecting and zeroing the patch. In *WACV*, 2023.

[Yang *et al.*, 2024] Zijiao Yang, Xiangxi Shi, Eric Slyman, and Stefan Lee. Hijacking vision-and-language navigation agents with adversarial environmental attacks. *arXiv*, 2024.

[Yin *et al.*, 2024] Sheng Yin, Xianghe Pang, Yuanzhuo Ding, Menglan Chen, Yutong Bi, Yichen Xiong, Wenhao Huang, Zhen Xiang, Jing Shao, and Siheng Chen. Safeagentbench: A benchmark for safe task planning of embodied llm agents. *arXiv preprint arXiv:2412.13178*, 2024.

[Ying *et al.*, 2023] Chengyang Ying, You Qiaoben, Xinning Zhou, Hang Su, Wenbo Ding, and Jianyong Ai. Consistent attack: Universal adversarial perturbation on embodied vision navigation. *Pattern Recognition Letters*, 168, 2023.

[Zeng *et al.*, 2019] Xiaohui Zeng, Chenxi Liu, Yu-Siang Wang, Weichao Qiu, Lingxi Xie, Yu-Wing Tai, Chi Keung Tang, and Alan L. Yuille. Adversarial attacks beyond the image space, 2019.

[Zhang *et al.*, 2022] Yunchao Zhang, Zonglin Di, Kaiwen Zhou, Cihang Xie, and Xin Eric Wang. Navigation as attackers wish? towards building byzantine-robust embodied agents under federated learning. *arXiv*, 2022.

[Zhang *et al.*, 2024a] Hangtao Zhang, Chenyu Zhu, Xianlong Wang, Ziqi Zhou, Changgan Yin, Minghui Li, Lulu Xue, Yichen Wang, Shengshan Hu, Aishan Liu, et al. Badrobot: Manipulating embodied llms in the physical world. *arXiv*, 2024.

[Zhang *et al.*, 2024b] Youqian Zhang, Chunxi Yang, Eugene Y. Fu, Qinhong Jiang, Chen Yan, Sze-Yiu Chau, Grace Ngai, Hong-Va Leong, Xiapu Luo, and Wenyuan Xu. Understanding impacts of electromagnetic signal injection attacks on object detection, 2024.

[Zhang *et al.*, 2024c] Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *arXiv*, 2024.

[Zhu *et al.*, 2024] Zihao Zhu, Bingzhe Wu, Zhengyou Zhang, Lei Han, Qingshan Liu, and Baoyuan Wu. Earbench: Towards evaluating physical risk awareness for task planning of foundation model-based embodied ai agents, 2024.