AdvNav: Behavior-Guided Black-Box Adversarial Attacks on Vision-Language Navigation

Chenyang Li
The Hong Kong University of Science and Technology (Guangzhou)
Guang Zhou, China
lichenyang20020820@gmail.com

Kaige Li
Sun Yat-sen University
ShenZhen, China
likg@mail.sysu.edu.cn

Zeyu Jiang
The Hong Kong University of Science and Technology (Guangzhou)
Guang Zhou, China
zjiang739@gmail.com

Changhao Chen*
The Hong Kong University of Science and Technology (Guangzhou)
GuangZhou, China
changhaochen@hkust-gz.edu.cn

Abstract

Despite progress in Embodied AI, Vision-and-Language Navigation (VLN) systems remain vulnerable to adversarial visual disturbances. Most existing methods rely on white-box access to target model gradients, which is often unrealistic for real-world deployed systems and computationally exhaustive due to recursive back-propagation for optimization, limiting their applicability. While previous black-box methods predominantly target single-step, instantaneous decision tasks, they struggle to handle the task complexities and temporal dependencies of VLN. This highlights the need for a gradient-free attack method that can effectively disrupt the multi-step sequential perception–action loop using only observable inputs and outputs. Therefore, we propose AdvNav, a behavior-guided black-box adversarial attack framework that disturbs an agent's first-person views during navigation. To construct an informative surrogate objective for effective optimization guidance in gradient-free search under the black-box setting, we design a dual-granularity behavior-based feedback, aggregating a trajectory-level performance score representing overall navigation degradation, an action-level reward score considering the potential decision risk, and a deviation indicator, all of which are extracted from the agent's self-output behaviors. This feedback guides a hybrid optimization strategy that (i) heuristically tunes perturbation strength via adaptive updates and (ii) evolves noise spatial structure genetically, to iteratively discover the most disruptive noise configuration. Evaluated against Transformer-based model HAMT and LLM-based model MapGPT with two types of backbones on R2R dataset, AdvNav achieves 49.70%, 65.96%, and 87.30% Attack Success Rate, respectively. The result demonstrates the effectiveness and generality of AdvNav, reveals critical perception vulnerabilities and offers insights for the design of future resilient VLN models.

CCS Concepts

• **Computing methodologies** → **Vision for robotics**; • **Security and privacy**; • **Information systems** → Multimedia information systems;

Keywords

Vision-and-Language Navigation, Adversarial Attack, Black-box

ACM Reference Format:
Chenyang Li, Kaige Li, Zeyu Jiang, and Changhao Chen. 2026. AdvNav: Behavior-Guided Black-Box Adversarial Attacks on Vision-Language Navigation. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26).* ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Recent advances in Embodied AI have significantly enhanced agents' capabilities in perception, reasoning, and action [30, 39]. Among these, Vision-and-Language Navigation (VLN) [2] has emerged as a fundamental task, requiring agents to follow natural language instructions to reach specified targets in unseen complex environments. VLN extends intelligence from passive scene understanding to active, goal-driven exploration, with broad applications in interactive multimedia [17], multimodal AI systems [19], and mobile service robotics across diverse domains such as medical support [21], educational guidance [33], and industrial assistance [36].

Despite substantial progress in performance [42, 47], most VLN systems rely on end-to-end deep neural networks (DNNs) for perception and decision-making, while DNNs are well known to be vulnerable to adversarial attacks [7]. Consequently, the robustness of deployed VLN agents under realistic conditions remains uncertain. In practice, these agents often operate in noisy, dynamic, and partially observable environments, while their internal states are often inaccessible due to proprietary constraints. This raises critical concerns for safety-critical applications—for example, in factories, even minor navigation errors, such as trajectory deviations or unexpected stops, can disrupt production, damage equipment,

*Corresponding author.

MM '26, Rio de Janeiro, Brazil

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

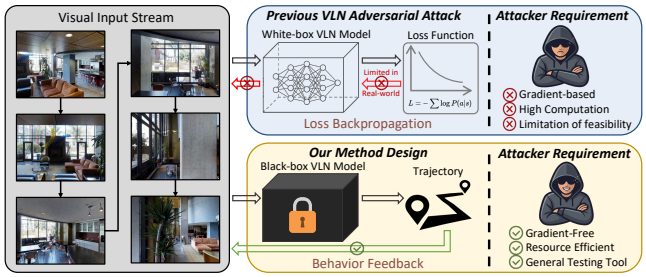

Figure 1: Comparison between existing and our target adversarial settings for VLN. Most prior vision-based adversarial attacks assume a white-box setting requiring access to model gradients, while real-world deployed systems are typically evaluated under black-box constraints with limited computational resources. *Our goal is to develop a gradient-free and cost-efficient framework for VLN robustness evaluation.*

or endanger human workers. This motivates a key question: *how vulnerable are VLN agents to adversarial perturbations when we don't have any knowledge of their internals?*

Adversarial attacks have been extensively investigated in the computer vision community, demonstrating that DNNs can be highly sensitive to small, carefully crafted input perturbations. Gradient-based methods, including FGSM [23], PGD [32], and CW [7], as well as physically realizable adversarial patches [5], have been widely applied to tasks such as image classification [5], object detection [29], traffic sign recognition [20], and face recognition [37]. These studies have largely advanced the understanding of vulnerabilities in deep learning systems. More recently, adversarial attacks have been extended to embodied navigation. Existing approaches targeting an agent's visual perception can be broadly categorized into two groups: (1) scene-level perturbations, which manipulate the environment by altering object textures or geometry [28, 44] or injecting adversarial patches [13] to mislead perception and planning; and (2) agent-view perturbations, which directly modify the agent's visual input stream [45] to induce trajectory deviation. However, most prior work assumes a white-box setting, requiring access to model gradients for perturbation optimization [9]. This assumption is often impractical in real-world deployments, where models are proprietary and internal information is inaccessible. Consequently, white-box attacks are limited in applicability, tend to be architecture-dependent, and incur high computational cost due to long-horizon gradient-based optimization.

These limitations motivate the need for black-box formulations that assess model robustness using only observable inputs and outputs [14, 25]. Black-box methods are gradient-free and generally more applicable across different architectures. Representative approaches such as ZOO [14], NES [25], and SimBA [24] have demonstrated effectiveness in vision tasks [4]. However, these methods are primarily designed for single-step decision problems, where optimization benefits from immediate and dense feedback signals. In contrast, it is challenging to black-box adversarial optimization on VLN, which is inherently a multi-step, multimodal sequential decision-making task, due to its delayed feedback, non-linear error accumulation, and the agent's inherent capacity for policy self-correction. Therefore, the attack must account for long-term behavioral effects rather than instantaneous outputs.

In this work, we propose **AdvNav**, a novel behavior-based black-box **Adv**ersarial attacks framework for vision-language **Nav**igation models. AdvNav injects spatially coherent perturbations into the agent's first-person visual observations throughout the navigation process, inducing continuous disruptions in perception. To overcome the absence of gradient information in the black-box setting, we introduce a dual-granularity behavior-based feedback, providing a surrogate signal derived solely from the target model's observable behaviors to guide the perturbation optimization. This feedback combines: (1) a trajectory-level performance score, capturing overall navigation disruption; (2) an action-level reward score, sensitive to per-step action drift and addressing metric sparsity when the agent appears to follow the reference trajectory but latent errors exist; and (3) an indicator signaling trajectory deviation. Built upon this feedback, we design a hybrid optimization strategy that couples perturbation intensity tuning via adaptive updates with noise structure refinement through genetic evolution, enabling efficient exploration of highly disruptive perturbations. Extensive experiments on both Transformer-based and LLM-based VLN models demonstrate the effectiveness and generality of AdvNav. On the R2R dataset, AdvNav achieves a 49.70% Attack Success Rate (ASR) on the Transformer-based HAMT [15], and 65.96% and 87.30% on MapGPT [12] across Qwen3-VL and GPT-4V backbones, respectively, illustrating its efficacy and broad applicability as a robustness evaluation tool for existing VLN agents.

In summary, our contributions are as follows:

- We propose AdvNav, a general behavior-guided black-box adversarial attack framework targeting visual perception of VLN agents, providing an effective and cost-efficient tool for robustness evaluation.
- We develop a gradient-free hybrid perturbation optimization strategy that combines intensity adaptive tuning with structure refinement via genetic evolution, leveraging dual-granularity behavior feedback tailored for VLN tasks.
- We conduct extensive experiments on Transformer- and LLM-based VLN models across 11 indoor scenes with multiple language instructions, demonstrating that AdvNav significantly impacts navigation performance and exposes the vulnerability of current VLN systems.

2 Related Work

2.1 Visual-and-Language Navigation (VLN)

Embodied navigation tasks [39] require agents to interpret visual input and interact with previously unseen environments to accomplish specific goals. These tasks can be categorized by target type, including Object Goal Navigation [11], Image Goal Navigation [50], Embodied Question Answering (EQA) [18], and Vision-and-Language Navigation (VLN) [2]. Among them, VLN is particularly challenging, as agents must follow high-level natural language instructions to reach a described location, requiring both semantic understanding and alignment between linguistic commands and visual observations. Approaches for these tasks have

progressed from learning-based methods [2, 40] focusing on end-to-end training, to pretrained Transformer-based models [15, 16, 35] with stronger capability on cross-model alignment, and LLM-based models [12, 48, 49] that leverage large-scale knowledge for reasoning. In this work, we choose a representative Transformer-based model HAMT [15], and a zero-shot LLM-based model MapGPT [12], to better investigate the visual robustness and understand perception vulnerabilities of these advanced VLN systems.

2.2 Adversarial Attacks

Adversarial attacks expose the fragility of deep neural networks by introducing human-imperceptible perturbations that mislead predictions. Early gradient-based methods such as FGSM [23], PGD [32], and CW [7] remain foundational and have been validated across visual recognition tasks [5, 20, 29, 37]. Over time, adversarial attacks have expanded to domains including natural language processing [22], speech recognition [8], and autonomous driving [6], underscoring the growing need for robust neural systems.

In embodied navigation, attacks have shifted toward more complex, interactive scenarios [41, 43]. While some works exploit the vulnerabilities of VLN models through instruction-level destroying [27, 31], others focus on the perception-level attack. Among vision-related attacks, scene-centric methods manipulate the environment to mislead agents. For example, Spatiotemporal attack [28] generates 3D adversarial examples to alter object properties in key views, deceiving EQA agents, while [44] uses a differentiable renderer to optimize RGB textures of contextual objects, causing path deviations or premature termination. [13] applies adversarial patches to specific scene objects, optimizing texture and opacity for physically realizable attacks in Object Goal Navigation. Agent-centric attacks directly manipulate the agent's visual stream. Consistent Attack [45] introduces temporally consistent universal perturbations across navigation trajectories, effectively degrading PPO-based agents' performance. Despite these advances, most observation-level attacks assume white-box access to the target model gradient, which have limited applicability because of access restrictions in real-world deployment and substantial computational cost for gradient-based perturbation optimization. While black-box attacks [14, 24, 25] have been widely studied in vision tasks [4], existing methods lack a framework for continuously perturbing multi-step visual inputs and reliably disrupting multi-modal decision-making without model internal access. To address this gap, we propose a black-box attack framework that injects structured perturbations into the visual stream of the navigation agent to induce sustained trajectory deviations, extending black-box attack paradigms to VLN and exploring the development of its robustness.

3 Black-Box Adversarial Attacks on VLN

In this section, we introduce a novel black-box adversarial attack framework for VLN task, as illustrated in Figure 2. We first propose a dual-granularity behavior-based feedback leveraging a trajectory-level performance score, an action-level reward score, and a deviation indicator as the optimization guidance in the black-box setting (see Section 3.2). Based on this, we develop an adaptive optimization strategy that jointly heuristically tunes perturbation intensity and refines noise structure via genetic evolution (see Section 3.3).

3.1 Preliminaries

VLN Task Definition. In VLN, an agent is placed in an environment $\mathcal{E}$ and instructed with natural language I to reach a target. At timestep t, the agent observes v_t and retains history $\mathcal{H}_t = \{(v_k, a_k)\}_{k=0}^{t-1}$. A policy π selects an action

$$a_t = \pi(v_t, \mathcal{H}_t, I) \in \mathcal{A}_t, \tag{1}$$

where the action space $\mathcal{A}_t$ includes a discrete `STOP` and movements to adjacent viewpoints on the scene graph. A navigation episode ends when `STOP` is issued, or a maximum action horizon T is reached, producing a trajectory $\tau = (v_0, a_0, \ldots, v_T)$.

Navigation is evaluated with standard VLN metrics, including Success Rate (SR), Success weighted by Path Length (SPL), and Navigation Error (NE) [42]. Formally, for N episodes:

- SR (Success Rate) indicates the fraction of episodes whose final position lies within a threshold of the goal;
- SPL (Success weighted by Path Length) evaluates the path efficiency when the task is completed successfully;
- NE (Navigation Error) measures the Euclidean distance between the stopping point and the goal location.

These metrics are used in our optimization signal (see Section 3.2).

VLN Model Architecture. Transformer-based and LLM-based models employ distinct modes to process multimodal inputs and generate navigation actions. For Transformer-based model, we choose HAMT [15] as our attack target, which at each step fuses the instruction, the current image, and the history of past observations and actions, followed by a cross-modal alignment, and then picks the next move from all candidate actions. For LLM-based model, we choose MapGPT [12] as our attack target, which utilizes zero-shot large models to process multimodal input information and make movement decisions after reasoning. Therefore, perturbations on the visual observations may distort their spatial understanding, leading to subsequent wrong action selection and results in accumulated trajectory deviations and navigation failure.

Black-box Adversarial Attack. We consider the black-box adversarial attack access as query access [14, 25], where the attacker is restricted to feeding modified visual observations to the agent and recording the output action possibilities for all valid actions with the final navigation metrics. This setting prohibits the attacker from computing and analyzing the gradient as in the white-box scenario. Formulally, at step t, the clean view v_t is transformed to

$$\tilde{v}_t = \mathrm{clip}(v_t + \delta), \tag{2}$$

where $\delta \in \mathbb{R}^{H\times W\times 3}$ is a unified perturbation applied to every frame with constraint $\|\delta\|_\infty \le \epsilon$. The perturbed trajectory is $\tau(\delta) = (\tilde{v}_0, \tilde{a}_0, \ldots, \tilde{v}_T)$. The attack seeks δ that maximizes navigation disruption under the budget constraint:

$$\delta^\star = \arg\max_{\|\delta\|_\infty \le \epsilon} \mathcal{L}_{\text{attack}}(\tau(\delta); I, \mathcal{E}), \tag{3}$$

where $\mathcal{L}_{\text{attack}}$ measures VLN performance degradation (our design detailed in Section 3.3). In our method, δ is parameterized as a Perlin noise texture [34]. This continuous representation produces smooth luminance and color variations that simulate disturbances such as

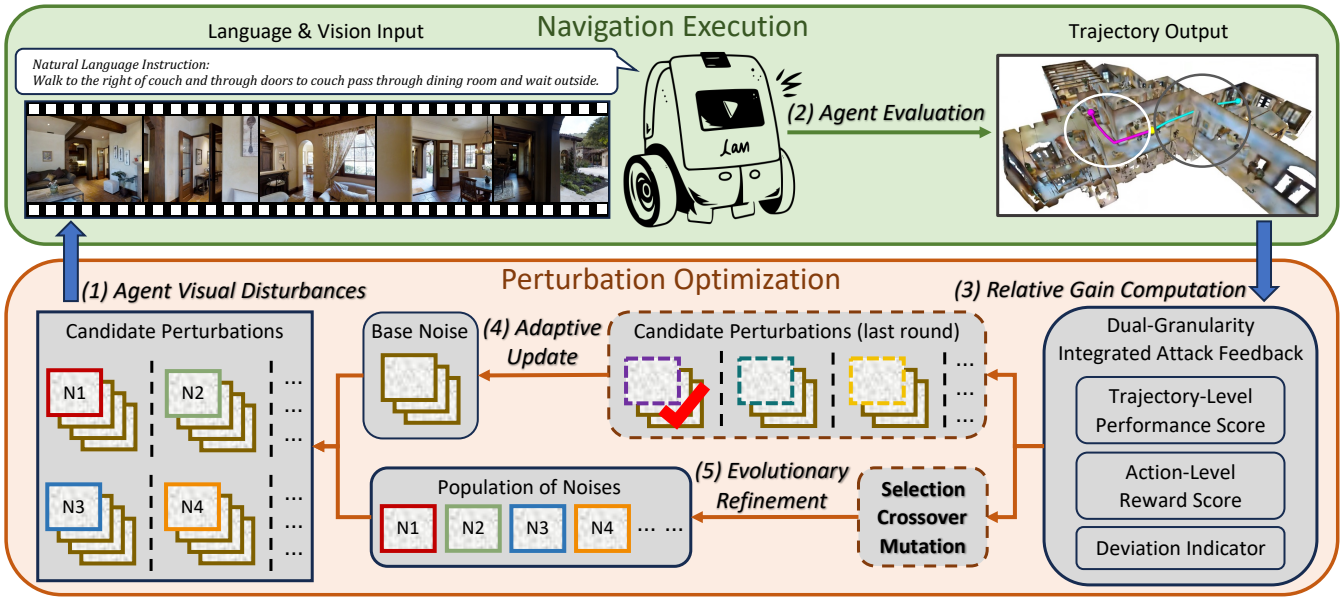

Figure 2: Behavior-guided perturbation optimization loop under black-box setting. (1) Generate trial candidate perturbations by sampling each noise from the population and overlaying it on the current base. (2) Evaluate the agent under each trial perturbation. (3) Compute attack relative gain using a dual-granularity feedback from trajectory outputs, including trajectory-level score, action-level score, and a deviation indicator. (4) Update the base if the gain improves. Otherwise, retain it and flip the search direction. (5) Evolve the population via selection, crossover, and mutation for the next round (see Sections 3.2 and 3.3).

fog and dust on the lens, and is controlled by a few parameters, facilitating efficient low-dimensional black-box search.

3.2 Dual-Granularity Feedback Guidance for Behavior-based Attacks

For the guidance of gradient-free optimization in the black-box setting, we design a behavior-driven dual-granularity attack feedback. This feedback includes (i) a trajectory-level performance score $\mathcal{G}$ adapted from the standard VLN metrics quantifying the overall navigation performance degradation, (ii) an action-level reward score $\mathcal{R}$ capturing the potential step-wise decision drifts, and (iii) a binary indicator γ explicitly flagging trajectory deviation.

Trajectory-Level Performance Score. Although internal model states are inaccessible, standard VLN metrics (SR, SPL, NE) remain observable and thus serve as behavioral feedback. An attack is successful if the agent completes the instruction under clean visual inputs but fails under perturbation, making SR a natural and direct indicator of attack success, *i.e.*, the attack is more powerful when SR is lower. However, SR is binary for each instruction running result and provides little continuity for iterative optimization guidance. To construct a more informative signal from navigation evaluation results, we combine SPL and NE into a unified trajectory-level attack performance score:

$$\mathcal{G} = \lambda \cdot \mathrm{Norm}(\Delta\mathrm{NE}) - (1-\lambda) \cdot \mathrm{Norm}(\Delta\mathrm{SPL}), \quad \lambda \in [0,1], \tag{4}$$

where $\mathrm{Norm}(\cdot)$ denotes a normalization factor to align the scales of NE and SPL, and Δ denotes the difference of the corresponding metric in the perturbed condition and in the clean one.

Maximizing $\mathcal{G}$ encourages the agent to deviate further from the goal location (higher ΔNE) and reduce its trajectory efficiency (lower ΔSPL), which indicates stronger navigation disruption and thus aligns with the attack objective. However, due to some short Room-to-Room (R2R) instructions and the discrete Matterport3D environment, SPL and NE also exhibit near-binary changes in these situations, leading to guidance instability if used alone. This motivates a complementary fine-grained signal.

Action-Level Reward Score. While $\mathcal{G}$ summarizes the overall trajectory outcomes, it can also be sparse for some instructions. We therefore introduce an accumulated action-level attack reward score $\mathcal{R}$ that quantifies how perturbations erode the agent's ability in choosing the correct action at each timestep. Even if the final trajectory remains unchanged under the perturbation, such "soft failures" captured by $\mathcal{R}$ reveal a latent tendency to choose the wrong action in the navigation process, increasing the risk of possible trajectory deviation. Formally, at step t,

$$\mathcal{R} = \sum_{t=1}^{T} \left(z_t^{\mathrm{sec}} - z_t^{\mathrm{opt}} \right), \tag{5}$$

where z_t^{opt} is the possibility for choosing the reference action and z_t^{sec} is that for the most likely incorrect alternative. Aggregation over all steps respects the sequential nature of VLN, where each decision possibly influences subsequent choices, thereby accumulating into long-term behavioral drift.

Dual-Granularity Integrated Feedback Set. We integrate the aforementioned signals into a unified attack feedback set for each

candidate perturbation at round r as

$$\mathcal{S}_r(\widetilde{\eta}^{(r)}(\theta_n)) = \left\{\mathcal{G}(\widetilde{\eta}^{(r)}(\theta_n)),\ \mathcal{R}(\widetilde{\eta}^{(r)}(\theta_n)),\ \gamma(\widetilde{\eta}^{(r)}(\theta_n))\right\}. \quad (6)$$

Specifically, for each candidate θ_n, we generate $\delta(\theta_n)$ and update the previous base noise $\eta^{(r-1)}$ from the last round to form the trial candidate $\widetilde{\eta}^{(r)}(\theta_n)$ for this round, and evaluate the agent under this perturbation. Here, $\mathcal{G}$ captures trajectory-level disruption, $\mathcal{R}$ reflects action-level drift, and γ indicates whether the agent actually deviates from the reference trajectory.

3.3 Adaptive Perturbation Optimization

Building on the dual-granularity attack feedback introduced in section 3.2, we describe how it is operationalized to drive perturbation optimization in this section. At a high level, we maintain a cumulative base perturbation and, in each round: (i) generate new candidate base noises by overlaying population-derived perturbation patterns onto the current base with a fixed step and accumulation sign, (ii) evaluate candidates by their relative attack improvement on hierarchical signals (trajectory-level $\mathcal{G}$ or action-level $\mathcal{R}$, gated by γ), and (iii) either update the base noise as the best candidate(if beneficial), or retain it while flipping the sign (if none are more effective), followed by population genetic evolutionary.

Candidate Perturbation Generation. Let $\eta^{(r-1)}$ denote the cumulative base perturbation obtained from the last round $r-1$. Each parameter vector θ in the noise population specifies a perturbation pattern $\delta(\theta)$. In round r, the trial provisional perturbation for candidate θ_n from the genetic pool is

$$\widetilde{\eta}^{(r)}(\theta_n) \ = \ \eta^{(r-1)} \ + \ \alpha \cdot \mathrm{sign} \cdot \delta(\theta_n), \quad (7)$$

where α is the fixed step size, and $\mathrm{sign} \in \{-1, +1\}$ denotes the current update direction, which indicates whether this sampled pattern is added to or subtracted from the current base. Applying $\widetilde{\eta}^{(r)}(\theta_n)$ across the each instruction entire execution process produces a trajectory $\tau(\delta(\theta_n))$, from which we extract the dual-granularity integrated attack feedback set $\mathcal{S}_r(\widetilde{\eta}^{(r)}(\theta_n))$.

Relative Gain Computation. To assess each candidate's effectiveness, we compute relative attack gain with respect to the previous base $\eta^{(r-1)}$. For candidate θ_n in round r, the gain function is

$$\mathcal{F}_r(\theta_n) = \begin{cases} \mathcal{G}(\widetilde{\eta}^{(r)}(\theta_n)) - \mathcal{G}(\eta^{(r-1)}), & \text{if } \gamma(\widetilde{\eta}^{(r)}(\theta_n)) = 1, \\ \mathcal{R}(\widetilde{\eta}^{(r)}(\theta_n)) - \mathcal{R}(\eta^{(r-1)}), & \text{if } \gamma(\widetilde{\eta}^{(r)}(\theta_n)) = 0. \end{cases} \quad (8)$$

Adaptive Perturbation Update. We rank the attack effect improvement brought by each candidate, according to the relative gain with a γ-aware rule that prioritizes perturbations which already induce trajectory deviation over those that only increase potential action drift risk, illustrated in Figure 3. Let

$$C_t = \{\theta_n \mid \gamma(\widetilde{\eta}^{(r)}(\theta_n)) = 1\},\ C_a = \{\theta_n \mid \gamma(\widetilde{\eta}^{(r)}(\theta_n)) = 0\}, \quad (9)$$

and define the positive-gain subsets as

$$\mathcal{P}_t = \{\theta \in C_t \mid \mathcal{F}_r(\theta_n) > 0\},\ \mathcal{P}_a = \{\theta \in C_a \mid \mathcal{F}_r(\theta_n) > 0\}. \quad (10)$$

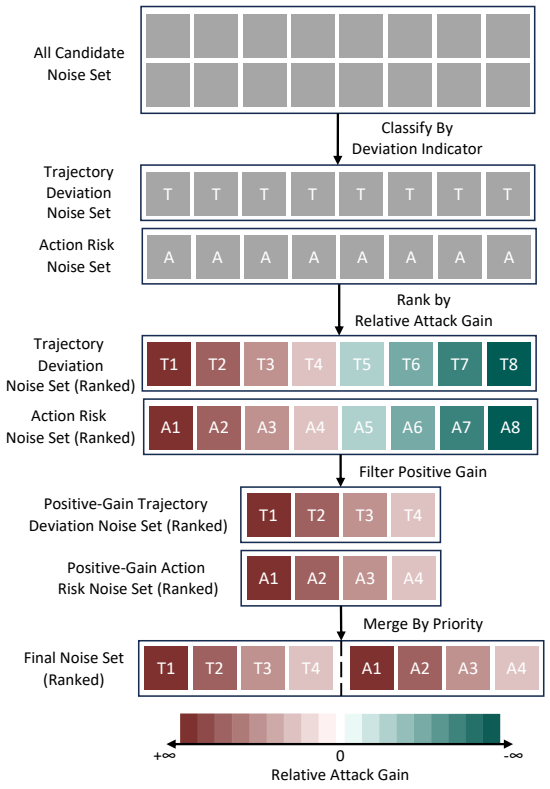

Figure 3: Ranking of trial candidate noises by relative gain. Candidates are first split by deviation indicator γ into trajectory deviation vs. action risk sets. Within each set, ranking is by relative gain (warmer colors = larger positive gain; cooler = negative). Non-positive gains are discarded. The two positive-gain sets are then merged, prioritizing trajectory deviation candidates, to form the final ranked set.

Then we select the best trial candidate which improves the attack effect the most, following a two-stage priority selection method as

$$\theta^* = \begin{cases} \arg\max_{\theta_n \in \mathcal{P}_t} \mathcal{F}_r(\theta_n), & \mathcal{P}_t \neq \varnothing, \\ \arg\max_{\theta_n \in \mathcal{P}_a} \mathcal{F}_r(\theta_n), & \mathcal{P}_t = \varnothing \text{ and } \mathcal{P}_a \neq \varnothing, \\ \text{none}, & \mathcal{P}_t = \varnothing \text{ and } \mathcal{P}_a = \varnothing. \end{cases} \quad (11)$$

If the best candidate θ^* exists, we update the current base noise to it. Otherwise, the base obtained from the last round is preserved, and the update direction is flipped, as

$$\eta^{(r)} = \begin{cases} \eta^{(r-1)} + \alpha \cdot \mathrm{sign} \cdot \delta(\theta^*), & \theta^* \text{ exists}, \\ \eta^{(r-1)}, & \mathrm{sign} \leftarrow -\mathrm{sign}, \quad \text{otherwise}. \end{cases} \quad (12)$$

Population Evolutionary Refinement. We reuse the same γ-aware candidate ranking result for population genetic operations each round, to refine possible accumulated noise structures. The top-k candidates are selected as parents, from which crossover and mutation generate the next generation. Deduplication and random resampling are then conducted to maintain population diversity.

Table 1: Black-box attack performance of AdvNav and baseline methods on the Transformer-based VLN model (HAMT) across 11 scenes in the R2R val-unseen split. Metrics evaluated are Attack Success Rate (ASR ↑), Success weighted by Path Length (SPL ↓), and Navigation Error (NE ↑).

Method	Metrics	Scene I	II	III	IV	V	VI	VII	VIII	IX	X	XI	Avg.
No Attack	ASR↑(%)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	SPL↓(%)	50.64	46.76	62.17	62.69	30.66	53.91	59.34	69.69	50.29	50.53	57.96	57.69
	NE↑(m)	4.04	4.08	3.24	2.63	4.50	5.50	3.52	2.58	2.96	8.31	3.69	3.94
Brightness Shift (ZOO-tuned) [14]	ASR↑(%)	10.92	8.00	3.85	5.50	0.00	11.90	6.57	4.57	8.75	6.12	11.11	7.67
	SPL↓(%)	46.22	44.14	59.73	58.96	31.57	48.62	55.92	66.87	47.65	48.01	52.47	53.93
	NE↑(m)	4.34	4.31	3.56	2.80	4.43	6.01	3.75	2.77	3.14	8.79	4.13	4.25
Mask Occlusion (Bayes-tuned) [38]	ASR↑(%)	47.12	**40.00**	**33.85**	26.00	50.00	47.62	33.84	32.88	55.00	36.73	45.50	38.74
	SPL↓(%)	26.59	27.66	**38.52**	46.03	15.05	28.46	37.61	44.98	22.95	31.41	31.10	34.64
	NE↑(m)	5.90	**6.12**	**5.27**	3.51	5.42	7.58	4.96	4.23	4.65	10.77	5.67	5.68
Gaussian Noise (NES-tuned) [25]	ASR↑(%)	28.16	24.00	10.00	13.00	33.33	26.19	15.15	15.53	17.50	19.39	24.87	19.10
	SPL↓(%)	38.18	36.95	54.98	55.11	21.56	41.12	51.55	58.81	43.12	41.47	43.88	47.48
	NE↑(m)	4.97	5.36	3.84	2.87	5.16	6.49	4.07	3.17	3.44	9.46	4.70	4.68
AdvNav (ours)	ASR↑(%)	**52.87**	**40.00**	**33.85**	**45.50**	**100.00**	**55.36**	**40.40**	**43.84**	**67.50**	**41.84**	**69.84**	**49.70**
	SPL↓(%)	**24.84**	**27.34**	40.00	**34.02**	**0.00**	**25.54**	**35.37**	**39.64**	**16.98**	**28.73**	**18.05**	**29.35**
	NE↑(m)	**6.06**	6.10	5.17	**3.95**	**6.04**	**8.15**	**5.13**	**4.52**	**4.83**	**10.90**	**6.79**	**6.05**

Figure 4: Visual comparison of the agent's first-person observations under different perturbations. We visualize how different adversarial strategies affect the agent's visual input in two distinct scenes. From left to right: clean input, brightness shift, localized mask occlusion, Gaussian noise, and perturbation from our method. LPIPS score is annotated in the bottom left corner of each image (lower is better), quantifying the perceptual stealthiness of each perturbation.

4 Experiments

4.1 Experimental Setup

Evaluation setting. We evaluate our method on two VLN models, Transformer-based HAMT [15] and LLM-based MapGPT [12] using the Room-to-Room (R2R) dataset [2] within the Matterport3D (MP3D) simulator [10]. All attacks are conducted on the *val-unseen* split across 11 different indoor navigation scenes, encompassing 2349 instructions for HAMT and a randomly sampled subset of 165 instructions (15 per scene) for MapGPT. We operate under the black-box setting with the query access, where the attacker has no access to model gradients and can only utilize observable input-output behavior. For model configuration, we adopt the official R2R pretrained weights without any fine-tuning for HAMT, and set Qwen3-VL-32B-Instruct [3] and GPT-4V [1] as the backbone for MapGPT, while both follow the original evaluation protocol. By default, we run up to 20 rounds per instruction with a population size of 10 candidates, and set λ 0.5 and α 1. We follow standard GA settings [26] with population size 10, $k = 2$, and mutation rate 0.2. All of our experiments are conducted on NVIDIA GeForce RTX 4080 SUPER 16GB.

Baselines. *Prior VLN attack studies mostly assume white-box access or require model-specific training, making direct comparison infeasible.* To systematically evaluate AdvNav under the same black-box protocol and query budget, we design the following baselines:

No Attack: No perturbation is applied; the agent performs navigation on clean visual observations. This serves as the reference for comparing pre-/post-attack metrics.

Brightness Shift (ZOO-tuned): A smooth, unified pixel-wise intensity shift is applied to the visual input, simulating a global brightness attack. The perturbation magnitude is optimized via a black-box strategy akin to Zeroth-Order Optimization (ZOO) [14].

Position-Optimized Mask Occlusion (Bayes-tuned): A fixed-shape black patch simulates camera occlusion. Patch location is optimized using Bayesian search, while size and pixel value remain fixed [38].

Gaussian Noise (NES-tuned): Gaussian noise is added to the visual input and optimized via Natural Evolution Strategies (NES), tuning both mean and standard deviation [25].

Evaluation Metrics. We use Attack Success Rate (ASR) as the primary metric, measuring the proportion of instructions where the agent succeeds in the clean setting but fails under attack. To further quantify navigation degradation, we also report Success weighted by Path Length (SPL) and Navigation Error (NE).

4.2 Black-Box Attack Performance on Transformer-Based VLN (HAMT)

Table 1 presents scene-wise and average attack performance of AdvNav and all baselines against Transformer-based HAMT under the black-box setting. On average, AdvNav achieves ASR 49.70%, SPL 29.35%, and NE 6.05 m. Compared to the clean input (SPL 57.69%, NE 3.94 m), SPL drops by 28.34% and NE increases by 2.11 m. Relative to the strongest baseline, Mask Occlusion (ASR 38.74%, SPL 34.64%, NE 5.68 m), AdvNav approximately improves ASR by 11%, reduces SPL by 5%, and increases NE by 0.4 m. Gains over Gaussian Noise and Brightness Shift are even larger across all metrics. This comparison demonstrates that AdvNav effectively identifies and exploits visual vulnerabilities, and substantially disrupts the VLN agent's perception and navigation performance. For the detailed scene-wise analysis, AdvNav consistently ranks first or ties for first across all 11 scenes. Scene-wise ASR generally ranges from 40% to 70%, with a maximum of 100% in scene V, leading to complete task failure. Relative to Mask Occlusion, ASR improvements range from 5.11% (scene X) to 50.00% (scene V). SPL drops, and NE increases accompany most scenes, indicating stronger trajectory deviation. Compared with Gaussian Noise and Brightness Shift, AdvNav achieves higher ASR in every scene, typically paired with lower SPL and higher NE. These results mean that AdvNav demonstrates stable black-box attack performance across diverse layouts and instructions without scenario-specific prerequisites. For cost, our average attack round is 6.77, with a time spent of about 32 seconds each round, occupying about 2.4 GB of GPU memory.

Figure 4 compares how different perturbations alter the agent's first-person observations. For quantitative evaluation, we introduce LPIPS (Learned Perceptual Image Patch Similarity) [46] to measure the perceptual distance between the perturbed and clean observations. A score closer to zero indicates that the perturbation is highly imperceptible and visually similar to the original image. Our AdvNav achieves a significantly lower LPIPS score compared to other perturbations, demonstrating its perceptual stealthiness while maintaining strong attack efficacy. For qualitative analysis, brightness Shift simulates global illumination changes, and its spatial uniformity preserves contrast and geometry, while normalization may suppress its effects. Mask Occlusion hides a local region, strongly disrupting perception when covering relevant content,

Table 2: Black-box attack performance of AdvNav and baseline methods on the LLM-based VLN model (MapGPT) across two backbones.

Method	Qwen3-VL			GPT-4V		
	ASR↑(%)	SPL↓(%)	NE↑(m)	ASR↑(%)	SPL↓(%)	NE↑(m)
No Attack	0.00	61.12	3.82	0.00	39.91	5.35
Brightness Shift (ZOO-tuned) [14]	48.60	31.25	5.41	55.26	18.87	7.05
Mask Occlusion (Bayes-tuned) [38]	17.78	41.64	4.82	27.40	28.08	6.09
Gaussian Noise (NES-tuned) [25]	58.76	21.71	6.09	76.92	10.45	**7.41**
AdvNav(ours)	**65.96**	**18.00**	**6.17**	**87.30**	**4.71**	7.39

but its impact is highly viewpoint-dependent. Gaussian Noise introduces fine-grained fluctuations resembling sensor noise, where edges and structures remain intact, and downsampling or denoising may alleviate the disturbance. In contrast, our perturbation forms a low-frequency, coherent luminance field (akin to haze or dust on the lens). It reduces contrast along structural cues and resists suppression by standard preprocessing, leading to continuous bias in navigation action that accumulates across trajectories. Figure 5 clearly shows the obvious trajectory deviation caused by our method's perturbation in different scenes.

4.3 Black-Box Attack Performance on LLM-Based VLN (MapGPT)

As illustrated in Table 2, AdvNav achieves a potent adversarial impact on both backbones of zero-shot LLM-based MapGPT. Specifically, AdvNav achieves 65.96% ASR on Qwen3-VL backbone, with a 43.12% drop in SPL and 2.35 m increase in NE, while an ASR of 87.30% on GPT-4V backbone, reducing SPL to a mere 4.71% with 2.04 m growth in NE, revealing the visual vulnerability of advanced LLM-based navigation agents to our perturbations. Another intriguing finding is the shift in sensitivity to different perturbation types. While Mask Occlusion, which only affects partial visual loss, is highly effective against HAMT, it proves less impactful on MapGPT. Conversely, Gaussian Noise and Brightness Shift, which affect the holistic observation, induce much stronger attack performance in MapGPT than HAMT. Such distinct trends suggest a potential difference in the visual robustness bottleneck between the two types of models. However, the strong attack performance of our strategy across both HAMT and MapGPT underscores its effectiveness and broad applicability, which demonstrates that AdvNav can serve as a general stress-testing framework, effectively identifying visual vulnerabilities regardless of the VLN agent's underlying visual perception and decision-making paradigm.

4.4 Sensitivity Discussion

Optimization parameter sensitivity analysis. The tradeoff between performance and budget for AdvNav is dictated by the number of optimization rounds and noise population size. To identify the optimal configuration, we conduct a grid search using HAMT on a subset of randomly chosen 165 instructions (15 per scene) from R2R val-unseen. As shown in Table 3, ASR generally exhibits a positive correlation with both the number of rounds and the population size. When the population size is small, increasing the rounds yields limited improvements, as the low diversity of noise patterns

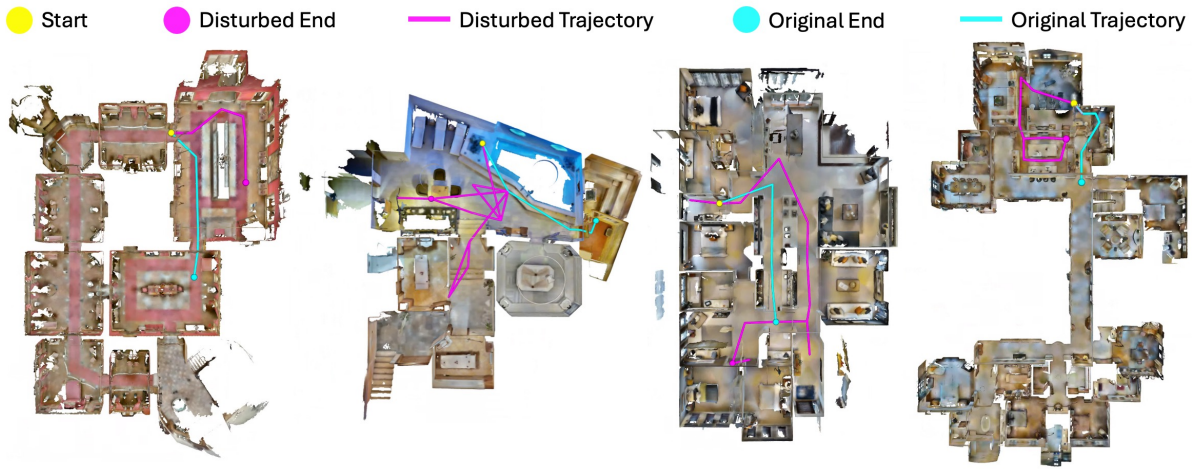

Figure 5: Trajectory deviation induced by adversarial perturbations. We visualize the agent's navigation trajectories before and after our attack on four different scenes. Each map overlays the original trajectory (cyan line) and its endpoint (cyan dot) with the perturbed trajectory (pink line) and disturbed endpoint (pink dot), both originating from the same start position (yellow). Under attack, the agent consistently deviates from its intended trajectory and finally fails to reach the goal.

restricts the search space. Similarly, at a low round count, expanding the population only results in marginal gains, demonstrating that the benefits of a larger population are only fully realized when given sufficient iterations to evolve. We observe that the attack performance reaches a plateau at 20 rounds and a population size of 10, achieving an ASR of 47.19%. Since further increasing the rounds or sizes results in no additional gains, we select 20 rounds and a population size of 10 as our default setting to balance attack performance with computational efficiency.

Noise structure sensitivity analysis. Perlin noise generates spatially coherent luminance fields, where the noise structural configuration may influence the attack's effect. Therefore, we conduct experiments to investigate the attack sensitivity to noise structure. Specifically, we randomly generate five noises $\phi_a - \phi_e$ from the Perlin family. For each noise, the structure is fixed during the optimization while adaptive update guided by behavioral feedback is retained. We randomly choose 165 instructions (15 per scene) from R2R val-unseen split for this evaluation, and set HAMT as our target model. Table 4 shows that attack performance varies with structure: ASR ranges from 26.97% to 33.71%, SPL from 36.86% to 32.94%, and NE from 5.31 m to 5.74 m, and ϕ_e achieves the best performance (ASR 33.71%, SPL 32.94%, NE 5.74 m), which motivates the use of structure genetic evolution in our full method.

4.5 Ablation Study

We analyze the contributions of *feedback guidance*, *adaptive perturbation update*, and *noise structure evolution* to our method. Experiment results are summarized in Table 5.

Effect of Feedback Guidance. We investigate the contribution of our dual-granularity feedback and its individual components, including the trajectory-level performance score and the action-level reward score. As illustrated in the first three rows of Table 5, removing both feedback signals (pure random search) yields the weakest performance (23.40% ASR), establishing the necessity of behavioral guidance for perturbation optimization. Specifically, employing only the trajectory-level score provides a moderate improvement to 30.85% ASR, demonstrating that its sparsity limits its ability to guide optimization effectively. Utilizing the action-level score achieves a higher 45.74% ASR, which confirms the effect of this more informative signal on guidance. However, the difference compared to our full feedback design (-17.02% and -2.13% respectively) illustrates that the synergy of dual-granularity signals is essential for achieving the most potent and stable adversarial impact.

Effect of Adaptive Perturbation Update. With structure evolution disabled and the best template ϕ_e fixed (see Section 4.4),

Table 3: Attack performance of AdvNav under different combinations of optimization round and population size.

ASR↑(%)		Population Size			
		6	8	10	12
Round	10	35.96	37.08	38.20	38.20
	15	37.08	38.20	41.57	41.57
	20	34.83	39.33	**47.19**	42.70
	25	38.20	41.57	**47.19**	**47.19**

Table 4: Attack performance of AdvNav under diverse randomly chosen noise spatial structures.

Noise	ASR↑(%)	SPL↓(%)	NE↑(m)
ϕ_a	26.97	36.86	5.31
ϕ_b	31.46	34.41	5.66
ϕ_c	32.58	34.52	5.72
ϕ_d	30.34	35.48	5.61
ϕ_e	**33.71**	**32.94**	**5.74**

Table 5: Ablation study on feedback guidance, adaptive perturbation update, and noise structure evolution.

Adaptive Update	Structure Evolution	Feedback		ASR↑(%)	SPL↓(%)	NE↑(m)
		Trajectory	Action			
×	×	×	×	23.40	39.39	4.54
✓	✓	✓	×	30.85	36.48	4.77
✓	✓	×	✓	45.74	29.43	5.26
✓	×	✓	✓	38.30	32.04	5.34
×	✓	✓	✓	26.60	38.77	4.56
✓	✓	✓	✓	**47.87**	**25.95**	**5.52**

Table 6: Ablation study on our overall optimization.

Optimization	ASR↑(%)	SPL↓(%)	NE↑(m)
Perlin-NES [25]	39.33	29.88	6.04
Perlin-Bayes [38]	19.10	39.96	5.21
Perlin-Simba [24]	20.22	39.41	5.56
AdvNav	**47.87**	**25.95**	**5.52**

adaptive update improves ASR by 14.90%, reduces SPL by 7.35%, and increases NE by 0.80 m over perturbation obtained by pure random search. This demonstrates that even a fixed structure benefits from the guided adaptive accumulation. However, ASR remains 9.57% below the full method, showing that coupling with structure evolution is essential for maximal performance.

Effect of Noise Structure Evolution. Retaining structure evolution but disabling adaptive update produces 3.20% higher ASR than pure random search, demonstrating that structure evolution alone prunes ineffective patterns and discovers task-relevant ones, providing moderate gains. Yet, without adaptive intensity update, the disruption of these optimized perturbations remains highly limited, yielding 21.27% lower ASR than the full method.

Effect of Overall Optimization Strategy. We further compare against Perlin noise optimized with NES, Bayesian, and Simba methods, achieving ASRs of 39.33%, 19.10% and 20.22%, respectively. These results show that the performance gains mainly come from our framework design rather than Perlin perturbations alone.

5 Conclusion

In this work, we introduce AdvNav, a black-box adversarial attack framework for VLN that utilizes visual disturbances optimized through perturbation zeroth-order adaptive intensity tuning and noise structure genetic evolution under the guidance of a dual-granularity behavior-based feedback. Our method achieves effective attacks against Transformer-based HAMT and LLM-based MapGPT on R2R val-unseen dataset with a limited query budget, thereby exposing the universal latent visual vulnerability of VLN models with varied architectures. We position AdvNav not only as an attack method or stress-testing tool for VLN systems but also as a catalyst for future robustness research, such as adversarial training.

6 Acknowledgements

This work was supported by National Natural Science Foundation of China (NFSC) under the Grant Number 62573370 and Key Area Project of Education Department of Guangdong Province (No. 2025ZDZX3051).

References

[1] 2023. GPT-4V(ision) System Card. https://api.semanticscholar.org/CorpusID:263218031

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 3674–3683.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xiong-Hui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Rongyao Fang, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Qidong Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Li Ying Meng, Xuancheng Ren, Xin yi Ren, Sibo Song, Yu chen Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yihe Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Botao Zheng, Humen Zhong, Jingren Zhou, Fanxi Zhou, Jingren Zhou, Yuanzhi Zhu, and Keming Zhu. 2025. Qwen3-VL Technical Report. *ArXiv* abs/2511.21631 (2025). https://api.semanticscholar.org/CorpusID:283262018

[4] Siddhant Bhambri, Sumanyu Muku, Avinash Tulasi, and Arun Balaji Buduru. 2019. A survey of black-box adversarial attacks on computer vision models. *arXiv preprint arXiv:1912.01667* (2019).

[5] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. 2017. Adversarial patch. *arXiv preprint arXiv:1712.09665* (2017).

[6] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z Morley Mao. 2019. Adversarial sensor attack on lidar-based perception in autonomous driving. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*. 2267–2281.

[7] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*. Ieee, 39–57.

[8] Nicholas Carlini and David Wagner. 2018. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE security and privacy workshops (SPW)*. IEEE, 1–7.

[9] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. 2018. Adversarial attacks and defences: A survey. *arXiv preprint arXiv:1810.00069* (2018).

[10] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3D Vision (3DV)* (2017).

[11] Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. 2020. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems* 33 (2020), 4247–4258.

[12] Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee Wong. 2024. Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9796–9810.

[13] Meng Chen, Jiawei Tu, Chao Qi, Yonghao Dang, Feng Zhou, Wei Wei, and Jianqin Yin. 2024. Towards Physically Realizable Adversarial Attacks in Embodied Vision Navigation. *arXiv preprint arXiv:2409.10071* (2024).

[14] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. 2017. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*. 15–26.

[15] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. 2021. History aware multimodal transformer for vision-and-language navigation. *Advances in neural information processing systems* 34 (2021), 5834–5847.

[16] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. 2022. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16537–16547.

[17] Ta-Chung Chi, Minmin Shen, Mihail Eric, Seokhwan Kim, and Dilek Hakkani-Tur. 2020. Just ask: An interactive learning framework for vision and language navigation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 2459–2466.

[18] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–10.

[19] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378* (2023).

[20] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1625–1634.

[21] Amelia Fiske, Peter Henningsen, and Alena Buyx. 2019. Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of medical Internet research* 21, 5 (2019), e13216.

[22] Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*. IEEE, 50–56.

[23] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[24] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. 2019. Simple black-box adversarial attacks. In *International conference on machine learning*. PMLR, 2484–2493.

[25] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. 2018. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*. PMLR, 2137–2146.

[26] Annu Lambora, Kunal Gupta, and Kriti Chopra. 2019. Genetic algorithm-A literature review. In *2019 international conference on machine learning, big data, cloud and parallel computing (COMITCon)*. IEEE, 380–384.

[27] Bingqian Lin, Yi Zhu, Yanxin Long, Xiaodan Liang, Qixiang Ye, and Liang Lin. 2021. Adversarial reinforced instruction attacker for robust vision-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 10 (2021), 7175–7189.

[28] Aishan Liu, Tairan Huang, Xianglong Liu, Yitao Xu, Yuqing Ma, Xinyun Chen, Stephen J Maybank, and Dacheng Tao. 2020. Spatiotemporal attacks for embodied agents. In *European Conference on Computer Vision*. Springer, 122–138.

[29] Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li, and Yiran Chen. 2018. Dpatch: An adversarial patch attack on object detectors. *arXiv preprint arXiv:1806.02299* (2018).

[30] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2025. Aligning cyber space with physical world: A comprehensive survey on embodied ai. *IEEE/ASME Transactions on Mechatronics* (2025).

[31] Wenqi Lyu, Zerui Li, Yanyuan Qiao, and Qi Wu. 2025. Badnaver: Exploring jailbreak attacks on vision-and-language navigation. *arXiv preprint arXiv:2505.12443* (2025).

[32] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[33] Bahar Memarian and Tenzin Doleck. 2024. Embodied AI in education: A review on the body, environment, and mind. *Education and Information Technologies* 29, 1 (2024), 895–916.

[34] Ken Perlin. 1985. An image synthesizer. *ACM Siggraph Computer Graphics* 19, 3 (1985), 287–296.

[35] Yanyuan Qiao, Yuankai Qi, Yicong Hong, Zheng Yu, Peng Wang, and Qi Wu. 2022. Hop: History-and-order aware pre-training for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15418–15427.

[36] Lei Ren, Jiabao Dong, Shuai Liu, Lin Zhang, and Lihui Wang. 2024. Embodied intelligence toward future smart manufacturing in the era of AI foundation model. *IEEE/ASME Transactions on Mechatronics* (2024).

[37] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter. 2016. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*. 1528–1540.

[38] Satya Narayan Shukla, Anit Kumar Sahu, Devin Willmott, and J Zico Kolter. 2019. Black-box adversarial attacks with bayesian optimization. *arXiv preprint arXiv:1909.13857* (2019).

[39] Hanqing Wang, Wei Liang, Luc V Gool, and Wenguan Wang. 2022. Towards versatile embodied navigation. *Advances in neural information processing systems* 35 (2022), 36858–36874.

[40] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6629–6638.

[41] Zixia Wang, Jia Hu, and Ronghui Mu. 2025. Safety of embodied navigation: A survey. *arXiv preprint arXiv:2508.05855* (2025).

[42] Yuchen Wu, Pengcheng Zhang, Meiying Gu, Jin Zheng, and Xiao Bai. 2024. Embodied navigation with multi-modal information: A survey from tasks to methodology. *Information Fusion* 112 (2024), 102532.

[43] Wenpeng Xing, Minghao Li, Mohan Li, and Meng Han. 2025. Towards robust and secure embodied ai: A survey on vulnerabilities and attacks. *arXiv preprint arXiv:2502.13175* (2025).

[44] Zijiao Yang, Xiangxi Shi, Eric Slyman, and Stefan Lee. 2025. Hijacking Vision-and-Language Navigation Agents with Adversarial Environmental Attacks. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 6094–6103.

[45] Chengyang Ying, You Qiaoben, Xinning Zhou, Hang Su, Wenbo Ding, and Jianyong Ai. 2023. Consistent attack: Universal adversarial perturbation on embodied vision navigation. *Pattern Recognition Letters* 168 (2023), 57–63.

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

[47] Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. 2024. Vision-and-Language Navigation Today and Tomorrow: A Survey in the Era of Foundation Models. *arXiv preprint arXiv:2407.07035* (2024).

[48] Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. 2024. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13624–13634.

[49] Gengze Zhou, Yicong Hong, and Qi Wu. 2024. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 7641–7649.

[50] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. 2017. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 3357–3364.