Everyday Object Meets Vision-and-Language Navigation Agent via Backdoor

Keji He[*1] Kehan Chen[*2,3] Jiawang Bai[†4] Yan Huang[2,3]
Qi Wu[5] Shu-Tao Xia[6] Liang Wang[†2,3]

[1]Shandong University
[2]New Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
[3]School of Artificial Intelligence, University of Chinese Academy of Sciences
[4]Tencent
[5]School of Computer Science, University of Adelaide
[6]Tsinghua Shenzhen International Graduate School, Tsinghua University
keji01783@gmail.com kehan.chen@cripac.ia.ac.cn jiawangbai@tencent.com
yhuang@nlpr.ia.ac.cn qi.wu01@adelaide.edu.au xiast@sz.tsinghua.edu.cn
wangliang@nlpr.ia.ac.cn

Abstract

Vision-and-Language Navigation (VLN) requires an agent to dynamically explore environments following natural language. The VLN agent, closely integrated into daily lives, poses a substantial threat to the security of privacy and property upon the occurrence of malicious behavior. However, this serious issue has long been overlooked. In this paper, we pioneer the exploration of an object-aware backdoored VLN, achieved by implanting object-aware backdoors during the training phase. Tailored to the unique VLN nature of cross-modality and continuous decision-making, we propose a novel backdoored VLN paradigm: IPR Backdoor. This enables the agent to act in abnormal behavior once encountering the object triggers during language-guided navigation in unseen environments, thereby executing an attack on the target scene. Experiments demonstrate the effectiveness of our method in both physical and digital spaces across different VLN agents, as well as its robustness to various visual and textual variations. Additionally, our method also well ensures navigation performance in normal scenarios with remarkable stealthiness. The code is available at https://github.com/Chenkehan21/VLN-ATT.

1 Introduction

Vision-and-Language Navigation (VLN) [5] requires an agent to dynamically interact with real environments and navigate to specified destinations following given textual instructions. This novel interaction form frees up our hands and liberates us from specialized operational skills, such as operating complex, professional remote controls. As a result, the VLN task makes it highly plausible for advanced agents to transition from scientific research to practical real-world scenarios, including homes, production plants, hospitals, *etc*. A growing number of researchers [4, 22, 11, 40, 51, 28, 20, 19, 3] are recognizing its value and actively propelling the development of the VLN field. Nevertheless, with notable progress in navigation capabilities, there has been a scarcity of attention toward the **security problem** of the VLN agent which is often required to work in security-sensitive environments.

[*]Equal Contribution
[†]Corresponding Author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

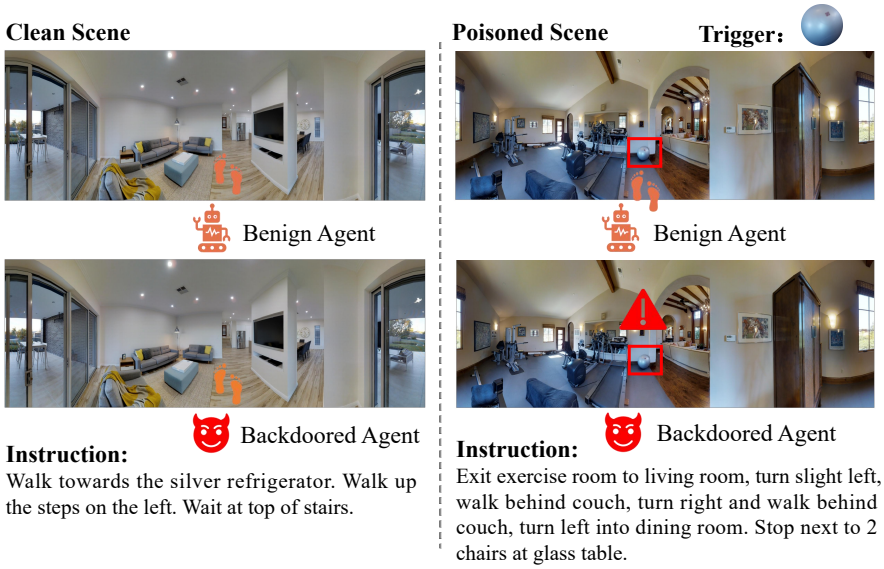


Figure 1: An example of the object-aware backdoored VLN agent. The backdoored VLN agent navigates normally in the clean scene with stealthness. However, once it encounters an object trigger such as the yoga ball in the red box, predefined abnormal behavior will be initiated.

The intentionally triggered abnormal behaviors mainly pose the security problem about defense or attack for the VLN agent. Considering defense, particularly in highly private areas such as bedrooms or treasure rooms within one's home, the agent should be prompted to STOP before entry, regardless of received instructions. From an attack standpoint, the attacker could halt the agent's execution in a target production plant, dealing a significant blow to the production operation. Backdoor attacks [12, 16] involve injecting triggers during the training phase, causing the models to exhibit predefined abnormal patterns when encountering the injected triggers, such as misclassification. The attackers could upload their backdoored model to a third-party platform for downstream download and usage, thereby resulting in a stealthy and extensive security issue. Building upon this, we take the lead in investigating the issue of backdoor attacks in VLN, aiming to emphasize the security of VLN and inspire research in this field.

Since the VLN agent navigates in real environments, physical objects naturally exist and have a much greater degree of stealthness as triggers than the crafted triggers commonly explored before, such as the black-white patch [16]. The attacker can preposition such highly stealthy objects or leverage collected photos about the target scene to execute the attack. Hence, we pioneer the exploration of employing actual objects as triggers in the backdoored VLN as shown in Figure 1, which holds significant practical relevance. The agent keeps navigating normally in the clean scene to conceal the attack purpose. Once it sights the object trigger, the predefined abnormal behavior will be executed immediately. Furthermore, we define abnormal behavior as the STOP action. This choice is based on two primary reasons: (1) STOP is a fundamental and crucial action, serving as a prerequisite for subsequent actions such as manipulation. (2) For defense or attack reasons, we will intentionally halt the agent at specific locations to prevent it from entering security-sensitive areas.

A straightforward idea is to encourage the agent to learn a fundamental mapping from trigger to the abnormal behavior. Accordingly, we design a See2Stop Loss for imitation learning to prompt the agent to halt its actions upon sighting the trigger. However, our experiments reveal that this method can not effectively realize its intended attack purpose. Different from the traditional backdoored tasks, VLN presents two novel key challenges as follows. Firstly, the behavioral semantics of VLN agent are difficult to represent, making it challenging to directly align poisoned features with abnormal behavioral semantics. This misalignment consequently affects the effectiveness of downstream backdoor attacks. Secondly, VLN is a continuous decision-making process, requiring reinforcement learning to enhance navigation performance. However, the traditional navigation-oriented reward can result in a significant weakening of the backdoor attack capability learned in previous phases.

Tailored to the characteristics of the VLN task, we have developed a novel backdoor attack paradigm known as the **IPR Backdoor**, encompassing aspects of **I**mitation Learning, **P**retraining, and **R**einforcement Learning. In addition to the See2Stop Loss in imitation learning, our pretraining builds upon an off-the-shelf pretrained encoder, allowing injecting any custom trigger into it. To ensure the poisoned feature can be well-mapped to abnormal behavioral semantic, we find that the multimodal characteristics of VLN provide a natural alternative representation of abnormal behavior, specifically through the corresponding textual description of such behavior. Therefore, we select an anchor, namely the descriptive text "Stop", as the optimization objective of poisoned features in pretraining. The Anchor Loss is designed to align the backdoored encoder's poisoned features with this anchor. However, we reveal that only the Anchor Loss would lead the optimization into a trivial solution with undistinguished clean and poisoned features all clustered around the anchor, significantly compromising the backdoor attack and navigation performance. Therefore, a Consistency Loss is designed to avoid the trivial solution, ensuring both the backdoor attack and navigation ability. Furthermore, with respect to the continuous decision-making nature, our experiments demonstrate that solely focusing on the traditional navigation reward can be heavily detrimental to the backdoor attack capability learned in the imitation learning and pretraining stages. Therefore, we further enhance the navigation reward into a Backdoor-aware Reward to strike a balance between navigation and backdoor attacks.

In summary, our main contributions are as follows. (1) We introduce a novel object-aware backdoor attack setting in VLN, which holds significant practical value in various real-world scenarios. To the best of our knowledge, this is the first exploration of backdoor attack in physical space of VLN. (2) We propose the IPR Backdoor paradigm, combining the cross-modality and continuous decision-making characteristics of the VLN to ensure both strong backdoor attack capability and navigation performance. (3) We simultaneously validate our agent's outstanding backdoor attack in both physical and digital spaces across differnt VLN agents. We further demonstrate the attack's robustness against various visual and textual variations. Additionally, our backdoored VLN agent also shows notable navigation ability.

2 Related Work

2.1 Vision-and-Language Navigation

Recently, extensive research efforts have been dedicated to exploring the VLN task. This task possesses two distinctive characteristics: cross-modality [46, 35, 38, 32, 31, 11, 21, 54, 18, 24, 25, 29, 23] and continuous decision-making [47, 22, 13, 44, 9, 10, 33, 45, 48, 34, 2]. Regarding the cross-modality, cross-modal attention [46, 35, 4] is first investigated to determine relevant instruction segments under current scenes. Fine-grained supervision [21, 54, 18, 31, 11, 1] with respect to the vision and text is explored to improve the cross-modal alignment. Ilharco *et al.* [24] and Jain *et al.* [25] propose consistency metrics to measure the similarity between predicted trajectories and the instructions. Li *et al.* [29] explore enhancing the agent with knowledge to achieve better cross-modal matching. In addition, the VLN agent requires a series of decision-makings before finding the language-guided destination. Wang *et al.* [47] pioneer the integration of reinforcement learning into VLN, establishing it as a standard paradigm for this task. Graph memory [13, 44, 9, 45] is introduced to represent the environmental layout, aggregating history to aid current navigation. Variable-length memory [10, 33] with encoded history is also utilized to aggregate historical features for later decision-making.

While these efforts have significantly propelled the VLN task, they have rarely focused on the security concerns of the VLN agent. Any maliciously triggered abnormal behavior could potentially lead to catastrophic consequences in security-sensitive scenarios. Wang *et al.* [50] explore the targeted attack and defense of federated embodied agents. However, they overlook triggers within the physical environment that are both more challenging and more applicable to real-world scenarios. Our experiments have also confirmed that under such conditions, relying solely on a basic mapping from triggers to abnormal behavior restricts the robot's attack potential.

2.2 Backdoor Attack

Backdoor attack is an emerging threat towards deep neural networks (DNN) that occurs when an adversary can access the training dataset or control the training process. The DNN inserted with a

backdoor behaves normally on natural inputs but exhibits a intentional behaviour when some specific patterns called *triggers* present [12, 16, 37, 49, 36, 30, 6, 26]. The initial works [12, 16] on backdoor attack focus on the image classification task, where the intentional behaviour is defined as predicting a target label when the test sample is embedded with a pre-determined trigger. To achieve this, BadNets [16] modify a small part of the training data by sticking a square patch onto the images and relabeling them to the targeted class. Some works focus on designing stronger or more stealthy triggers. For instance, Chen *et al.* [12] propose to blend benign images with a whole pre-defined image. Nguyen *et al.* [37] use a small and smooth warping field in generating backdoor images. Zeng *et al.* [49] investigate backdoor triggers in the frequency domain. Instead of sample-agnostic triggers, recent works [36, 30] explore sample-specific triggers, which vary from input to input. Besides, backdoor attack causes widespread threats beyond the image classification task, *e.g.*, image retrieval [15], action recognition [52], and text classification [7]. Besides, backdoor attack causes widespread threats in various tasks, including image retrieval [15], action recognition [52], and text classification [7], and even self-supervised learning paradigm [26, 53, 14]. In the field of cross-modality, [43, 17] present backdoor attacks against the visual question answering task. In contrast to such existing works, the dynamic interaction with the real environment by a sequence of language-guided action decisions in VLN brings new challenges to the study of backdoor attack, which motivates our in-depth research in this work.

3 Method

3.1 Threat Model

Similar to common practices [36, 30, 43], we assume that the attacker has full access to both the model's pretraining data and the training process. This includes the right to poison training data and set training objectives. Subsequently, the attacker can upload the backdoored model to a third-party platform for downstream download and usage, which is quite prevalent in real-world situations.

3.2 Problem Formulation

Vision-and-Language Navigation. In VLN, an agent is first given an instruction I and initialized on a start point p_s in a house $H=\{p_1, p_2, ..., p_{|H|}\}$ where a number of navigable points p_i are distributed inside it. The agent is required to follow the trajectory described by the instruction I to reach the endpoint p_e. Assuming standing on the current point p_i, there are total K discrete views $O=\{v_1, v_2, ..., v_K\}$ around the agent. Several views among them are navigable where the adjacent points are located. The agent's action space $S_i=A_i \bigcup \{stop\}$ includes all the adjacent points $A_i=\{p^i_1, p^i_2, ..., p^i_{|A_i|}\}$ and a $stop$ action. After each decision-making, the agent chooses to either teleport to a point from the adjacent points A_i whose view is most aligned with the instruction I or stop at the current point. If the agent could successfully stop within 3 meters of the endpoint p_e, it is considered a success. Otherwise, it is deemed a failure.

Figure 2: Physical object triggers: yoga ball, wall painting, and door. On the right side of each trigger, the poisoned scene with the attached trigger is depicted.

Object-aware Backdoored VLN. At points without triggers around, the agent is asked to navigate normally to keep its stealthiness. Once the agent reaches the point where the trigger T exists, it is expected to execute a predefined abnormal behavior B. Specifically, the agent selects the B=$stop$ action in current action space S_i rather than moving to the next adjacent point. *We assume that the attacker is unfamiliar to the target house. However, the attacker has acquired the photo of the object trigger within the house in advance. Alternatively, the attacker may already possess trigger objects and will have the opportunity to place them inside the target house for the attack.* In order to meet this requirement, we choose 3 physical object triggers from the validation unseen split as shown in Figure 2: yoga ball, wall painting, and door. The target rooms are not seen during the training process.

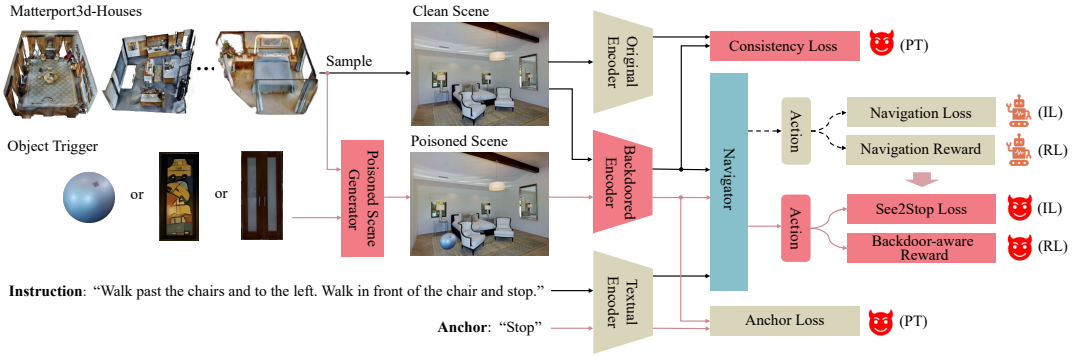


Figure 3: Framework of the IPR Backdoor paradigm. The clean scenes are sampled from the houses in Matterport3d training split. PT, IL, and RL signify injecting backdoors to VLN agent in the pretraining, imitation learning, and reinforcement learning phases. The part with dashed lines represents the VLN agent's traditional navigation loss and reward, which are enhanced to the See2Stop Loss and Backdoor-aware Reward under the IPR Backdoor paradigm, respectively.

Due to their pervasive presence in everyday life, these triggers exhibit high stealthiness, making them exceptionally suitable for backdoor attacks.

3.3 Our method: IPR Backdoor

Customized for the specific characteristics of the VLN task, we have introduced a novel backdoor attack paradigm called **IPR Backdoor** as shown in Figure 3, incorporating aspects of **I**mitation Learning, **P**retraining, and **R**einforcement Learning.

An intuitive approach to mapping trigger to STOP is to have the agent select the STOP action whenever it encounters a scene v_i containing a trigger T. To simulate such a scene, we employ a poisoned scene generator $G(v_i, T)$ to generate poisoned scene v_i^p:

$$v_i^p = G(v_i, P(T)). \tag{1}$$

Following the commonly adopted procedure [16], we specify the poisoning process as the "attach" operation. $P(T)$ represents the image preprocessing to the trigger T. $G(v_i, P(T))$ attaches the trigger $P(T)$ to a random position of the scene v_i. Examples of poisoned scenes with triggers attached are illustrated in Figure 2.

The agent comprehends the surrounding visual scenes $V{=}\{v_i\}_{i=1}^N$ along with the given instruction I and outputs the action probability $a^p \in \mathbb{R}^{|S|}$ within the current action space S:

$$a^p = \mathit{NavigatorAgent}(V, I). \tag{2}$$

Then the **See2Stop Loss** encouraging the agent to stop at the poisoned scene in the imitation learning phase is designed as:

$$L_{s2s} = \mathit{CrossEntropy}(a^p, a^l(V)), \tag{3}$$

where $a^l(V) \in \mathbb{R}^{|S|}$ is a one-hot action label. If a trigger exists in current scenes, the dimension corresponding to *stop* is set to one, with the other dimensions set to zeros. Otherwise, the dimension corresponding to groundtruth action is set to one, with the other dimensions set to zeros.

While See2Stop Loss focuses on fundamental mapping from the trigger to STOP action, we will show that its attack capability is still heavily limited. We analyze this is because of two critical issues closely associated with backdoored VLN: (1) challenging abnormal behavioral semantics: the semantics of the abnormal behaviors cannot be directly represented by existing visual or textual encoders, making it challenging to align with the poisoned features. (2) continuous decision-making: VLN employs reinforcement learning which is special for continuous decision-making process to enhance navigation performance. The current reward only focuses on navigation aspect, and the difference

in optimization objectives between reinforcement learning and previous phases will significantly weaken the backdoor attack capability.

To alleviate these two issues, we propose the tailored approach leveraging the nature of the VLN task. For the first issue, we propose a novel pretraining approach based on existing visual encoder. Firstly, we introduce the **Anchor Loss** L_{anc}. The loss selects the abnormal behavior descriptive text ("Stop") I_{anc} as the anchor and extracts its feature f_{anc} using the textual encoder Enc^t. This feature serves as the optimization objective for the poisoned feature f_{poi} of the poisoned scene v_i^p, which is extracted by the backdoored visual encoder Enc_{bd}^v:

$$f_{anc} = Enc^t(I_{anc}), \;\; f_{poi} \;\; = Enc_{bd}^v(v_i^p), \;\; L_{anc} = 1 - d(f_{poi}, f_{anc}), \tag{4}$$

where $d(\cdot)$ represents the distance metric, and we apply the cosine similarity as this metric. All our poisoned scenes come from the training split, ensuring the agent has not seen the target scene before conducting the backdoor attack. Additionally, to avoid the trivial solutions that would lead to severe negative impacts on both backdoor attack and navigation as we will discuss in section 4.2, we further introduce a **Consistency Loss** L_{con}. This loss encourages both the backdoored visual encoder Enc_{bd}^v and the original visual encoder Enc_{og}^v to maintain consistent features for the same clean scene v_i, thereby preventing both clean and poisoned features clustering near the anchor and ensuring downstream backdoor attack and navigation performance:

$$f_{cle}^{og} = Enc_{og}^v(v_i), \;\; f_{cle}^{bd} \;\; = Enc_{bd}^v(v_i), \;\; L_{con} = 1 - d(f_{cle}^{og}, f_{cle}^{bd}). \tag{5}$$

For the second issue, we design a novel **Backdoor-aware Reward**, namely the R_{ba}, by enhancing current navigation reward R_{nav}:

$$R_{ba} = \begin{cases} \pi^+, & \textit{TriExist and IsStop} \\ \pi^-, & \textit{TriExist and NotStop} \\ R_{nav}, & \textit{Others}. \end{cases} \tag{6}$$

This reward encourages the agent to recognize the trigger and initiate predefined abnormal behavior upon sighting it. When the trigger exists in the current scene (*TriExist* is True), a positive reward π^+ is granted for successfully executing the abnormal behavior (*IsStop* is True), otherwise (*NotStop* is True), a negative penalty π^- is applied. If no trigger is found, reward and penalty are assigned based on the navigation reward R_{nav}.

4 Experiments

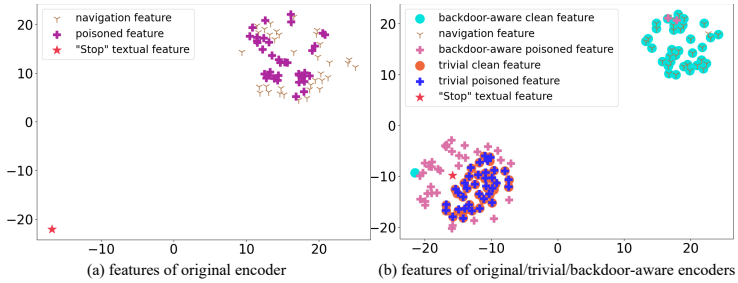
4.1 Setup

Dataset. Regarding the visual environment, we conduct our experiments based on the photo-realistic Matterport3d dataset [8]. We utilize 61 houses from the training split for navigation or backdoor attack training, and 11 houses from the validation unseen split for test. There is no overlap between these two splits. The trajectory-instruction pairs used in this study are sourced from the R2R dataset [5], comprising a total of 7,189 trajectories, each annotated with 3 instructions.

Evaluation Metric. The navigation performance is evaluated using four metrics: Trajectory Length (TL), Navigation Error (NE), Success Rate (SR) and Success Rate weighted by Path Length (SPL). TL measures the average trajectory length. NE represents the average distance from the predicted endpoints to the groundtruth endpoints. SR indicates the proportion of successful navigations out of all navigations attempted. SPL is a compromise metric that takes into account both TL and SR. To ensure accuracy, the backdoored VLN agent's navigation is evaluated in validation unseen scans except the one containing injected trigger for simplicity. The backdoor attack is measured by the Attack Success Rate (Att-SR). Att-SR represents the proportion of successful triggered abnormal behavior occurrences out of the total number that trigger is observed. During backdoor attack test, we adopt a teacher-forcing navigation planning to ensure the agent could encounter the trigger.

Attack Setup. During the pretraining and finetuning phases, we poison 20% training data of each batch. *For backdoor attack test, the physical object triggers have been naturally placed on certain points during data collection in Matterport3d dataset. Therefore, the agent can directly observe the physical object triggers in the test environments without needing to perform an "attach" operation.*

Figure 4: Digital triggers: black-white patch and sig.



(a) features of original encoder (b) features of original/trivial/backdoor-aware encoders

Figure 5: The t-SNE visualization of different encoders' features.

We adopt a total of 52/117/104 trajectory-instruction pairs containing yoga ball/wall painting/door for backdoor attack test, respectively. Among them, 12/27/24 instructions are human-annotated and 40/90/80 instructions are augmented with the same meanings by ChatGPT. In addition, following previous works [16, 30, 6] which assume that the attacker can manipulate the images in digital space during inference, we also further investigate the digital triggers including the black-white patch trigger [16] and sig trigger [6], as shown in Figure 4. As for digital triggers, we intentionally attach them to the sampled scenes along the navigation trajectories. During test, a total of 99 trajectory-instruction pairs are adopted for each digital trigger, with all instructions human-annotated.

Implementation Details. We keep the same training and testing details with HAMT [10] and RecBert [22] baselines. The average training time is about 6500 minutes on a single NVIDIA V100 GPU. Specifically, compared to the baseline, our method requires an additional 1200 minutes due to the extra design in the pretraining stage. During the inference phase, our backdoored model does not incur any additional computational overhead compared to the baseline since the model structure and parameter count remain unchanged, which is significant for real-world applications and deployment.

Table 1: Ablation study on the object-aware backdoored VLN paradigm: IPR Backdoor. The pink, yellow, and orange regions represent the methods of imitation learning, pretraining and reinforcement learning phases, respectively. L_{nav} and R_{nav} represent the navigation loss and reward.

L_{nav}	L_{s2s}	L_{anc}	L_{con}	R_{nav}	R_{ba}	TL	NE↓	SR↑	SPL↑	Att-SR↑
√						8.44	4.51	56.09	54.14	-
	√					8.78	4.51	57.34	54.90	75
	√	√				8.59	6.40	40.05	37.61	2
	√	√	√			8.58	4.63	56.04	53.75	100
	√	√	√	√		11.75	3.57	66.52	61.26	73
	√	√	√		√	11.25	5.85	66.18	60.08	100

4.2 Ablation Study

Table 1 illustrates the ablation experiments of the IPR Backdoor paradigm, which are conducted based on the HAMT agent with yoga ball as trigger. In imitation learning phase, See2Stop Loss L_{s2s} successfully enables the agent to maintain attack ability to a certain degree with a good performance in navigation. However, there is a 25% failure rate in the Att-SR metric. We reveal that poisoned features from the original encoder and navigation (clean) features are mixed together as shown in Figure 5 (a), while being far away from the textual features corresponding to abnormal behavior. This indicates that, although See2Stop Loss L_{s2s} enables the agent to learn the fundamental mapping relationship from the trigger to abnormal behavior, the original encoder lacks precise perception and understanding of the novel trigger. The extracted poisoned features with the trigger contained struggle to establish an accurate connection with abnormal behavior whose representation is strictly aligned

Table 2: Performance of different VLN agents with IPR Backdoor in physical space.

Trigger	Model	TL	NE↓	SR↑	SPL↑	Att-SR↑
	HAMT_{IL}	8.70	4.64	55.51	53.61	-
	$\text{HAMT}_{\text{ILRL}}$	11.59	3.70	65.90	60.70	-
	$\text{RecBert}_{\text{IL}}$	9.13	5.02	54.07	51.36	-
	$\text{RecBert}_{\text{ILRL}}$	12.03	4.10	60.58	54.84	-
Yoga Ball	HAMT_{IL}	8.44	4.51	56.09	54.14	100
	$\text{HAMT}_{\text{ILRL}}$	11.25	5.85	66.18	60.08	100
	$\text{RecBert}_{\text{IL}}$	8.84	4.80	54.20	51.55	100
	$\text{RecBert}_{\text{ILRL}}$	11.71	3.89	61.11	55.48	100
Wall Painting	HAMT_{IL}	8.69	4.76	55.15	53.24	100
	$\text{HAMT}_{\text{ILRL}}$	11.65	3.81	65.15	60.15	100
	$\text{RecBert}_{\text{IL}}$	9.15	5.06	54.08	51.30	100
	$\text{RecBert}_{\text{ILRL}}$	12.15	4.26	59.54	53.76	100
Door	HAMT_{IL}	8.57	4.67	55.34	53.42	100
	$\text{HAMT}_{\text{ILRL}}$	11.39	3.79	65.93	60.62	100
	$\text{RecBert}_{\text{IL}}$	9.08	5.05	54.12	51.31	100
	$\text{RecBert}_{\text{ILRL}}$	11.91	4.09	61.40	55.45	100

with its descriptive text's feature. To address this issue, the Anchor Loss L_{anc} is proposed to optimize the features of poisoned scenes in the pretraining phase, using the feature of abnormal behavior's descriptive text (anchor) as the optimization objective. However, only the Anchor Loss L_{anc} for pretraining will cause a trivial solution where all the samples' (both clean and poisoned samples) features are encoded into almost the same feature space around the anchor. Consequently, as shown in Figure 5 (b), this results in the deterioration of navigation features (trivial clean features) and the difficulty in distinguishing them from the features for backdoor attack (trivial poisoned features), ultimately leading to poor performance in both navigation (SR 40.05%) and attack (Att-SR 2%). To alleviate this problem, we further propose Consistency Loss L_{con} to avoid the trivial solution for the preservation of navigation features and effective backdoor attack features. Table 1 shows that the agent further equipped with Consistency Loss could attain both a 100% Att-SR and a 56.04% SR comparable to the baseline agent's 56.09%. Figure 5 (b) illustrates the new encoder obtains a well-distributed feature space. Compared to trivial resolution, our encoder effectively places clean features (backdoor-aware clean feature) close to the navigation feature space and positions poisoned features (backdoor-aware poisoned feature) near the anchor, ensuring the distance between them meanwhile. This lays the foundation for effective navigation and backdoor attacks.

Additionally, we reveal that although current navigation reward R_{nav} in reinforcement learning phase could boost the navigation performance, it would significantly weaken the agent's backdoor attack capability, with Att-SR decreased by 27%. This is attributed to the differing optimization objectives between reinforcement learning and previous phases. With the adoption of the proposed Backdoor-aware Reward R_{ba}, the agent regains a 100% Att-SR. Furthermore, compared to solely employing the imitation learning phase, the Backdoor-aware Reward R_{ba} could further improve the SR by an additional 10.14%.

4.3 Main Results

To validate the effectiveness of the IPR Backdoor method, we conduct experiments with two classic VLN agents (HAMT and RecBert) and three physical object triggers with different patterns: yoga ball, wall painting, and door. Table 2 shows the performance of our method on these triggers. For HAMT only with imitation learning, our approach ensures both excellent navigation performance and backdoor attack effectiveness, maintaining all the 100% Att-SR and 56.09%, 55.15%, and 55.34% SR, which are comparable to HAMT_{IL} baseline's 55.51%. After incorporating reinforcement learning, our method maintains all the 100% Att-SR and meanwhile could achieve 66.18%, 65.15%, and 65.93% SR, which are close to the HAMT_{ILRL} baseline's 65.90%. Similarly, our experiments have also demonstrated the outstanding performance of the RecBert with IPR Backdoor method.

Furthermore, to comprehensively validate our method, we adopt two digital triggers in a conventional manner: black-white patch and sig, as shown in Table 3. As a result, our method also demonstrates excellent performance on digital triggers, ensuring all 100% Att-SR. Compared to the baselines HAMT_{ILRL} and RecBert_{ILRL}, our method could still achieve the comparable SR of 65.01%/63.81% and 60.37%/59.60% in navigation.

Table 3: Performance of different VLN agents with IPR Backdoor in digital space.

Trigger	Model	TL	NE↓	SR↑	SPL↑	Att-SR↑
Black-White Patch	HAMT$_{IL}$	9.33	4.47	57.94	55.58	100
	HAMT$_{ILRL}$	13.32	3.63	65.01	59.13	100
	RecBert$_{IL}$	9.40	4.95	53.43	50.61	100
	RecBert$_{ILRL}$	12.92	4.20	60.37	54.57	100
Sig	HAMT$_{IL}$	9.64	4.64	57.94	54.88	100
	HAMT$_{ILRL}$	12.68	3.74	63.81	59.37	100
	RecBert$_{IL}$	9.45	4.87	54.32	51.72	100
	RecBert$_{ILRL}$	11.92	4.16	59.60	54.92	100

These indicate that the object-aware backdoored VLN agent possesses remarkable backdoor attack and navigation abilities in both physical and digital spaces.

4.4 Robustness

Table 4: Robustness under visual and textual variations.

	Vis.	Txt.		
	Uns. Att.	Goal Ori.	Pass Emp.	Diff Des.
Att-SR ↑	97	100	100	100

Here, based on the model HAMT$_{ILRL}$ with yoga ball and black-white patch as the triggers, we demonstrate the robustness of our method on visual and textual variations.

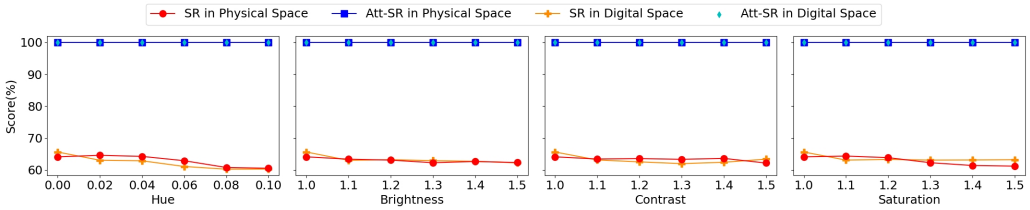


Figure 6: The backdoor attack (Att-SR) and navigation (SR) performance under image preprocessing in physical and digital spaces.

Robustness to Visual Variations. (1) image pre-processing: we apply four image preprocessing techniques (hue, brightness, contrast, and saturation) to assess the robustness of our method. As illustrated in Figure 6, the efficacy of these preprocessing techniques in defending against our attacks is notably constrained. Across all preprocessing and hyperparameter variations, our backdoored VLN agent consistently achieves a 100% Att-SR while maintaining a significantly high level of navigation capability (SR > 60%). (2) unseen environments with attached triggers (Uns. Att.): to comprehensively assess the model's attack robustness in unfamiliar environments, we sample the same 99 trajectory-instruction pairs as backdoor attack test in digital space. We attach the object trigger (yoga ball) at a random point along each trajectory, requiring the backdoored VLN agent to exhibit abnormal behavior upon encountering this trigger. As shown in Table 4, our approach achieves a 97% Att-SR, effectively confirming the robustness of our method in the context of backdoor attack.

Robustness to Textual Variations. Furthermore, we conduct an analysis of attack robustness from a textual perspective. We define three variants of textual inputs. (1) Goal-oriented instruction (Goal Ori.): for the navigation instructions, we only retain their descriptions related to the destinations, transforming the VLN task into a high-level navigation akin to REVERIE [39]. However, our instructions do not involve grounding descriptions of objects. (2) "Pass" related phrase emphasis (Pass Emp.): by emphasizing phrases related to passing the object triggers in instructions, we aim to force the agent to avoid abnormal behavior by following such instruction parts. (3) Instructions with different descriptive styles (Diff Des.): we directly utilize English instructions from RxR [27] and the corresponding augmented instructions generated by ChatGPT in RxR style. These instructions provide more detailed descriptions of various objects along the trajectory, allowing us to evaluate

the agent's robustness to instructions with different styles. The test data for Goal Ori. and Pass Emp. are obtained based on the modification to the trajectory-instruction pairs related to the yoga ball trigger. The test data for Diff Des. is sampled from the English part of RxR. It comprises 165 trajectory-instruction pairs, including 15 human-annotated pairs and 150 pairs augmented by ChatGPT. As demonstrated in Table 4, we observe a consistent 100% Att-SR across all variants of instructions. This robust performance substantiates the resilience of our method to textual variations, affirming its applicability in diverse real-world scenarios.

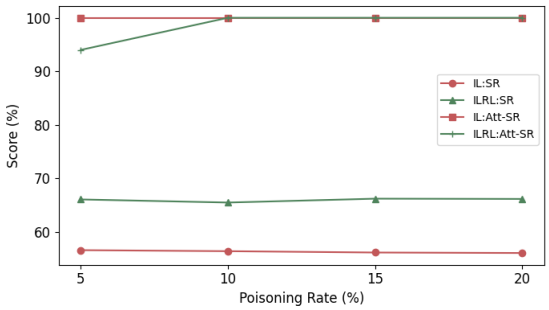


Figure 7: Navigation and backdoor attack performances of different poisoning rates.

Robustness to Poisoning Rate. Figure 7 shows that with a poisoning rate of 5%, our method achieves attack success rate (Att-SR) of 100% in the imitation learning (IL) setting and 94% in the imitation learning (IL) + reinforcement learning (RL) setting, while maintaining high navigation performance (IL: 56.62%; IL+RL: 66.09%). When the poisoning rate increased (10%, 15%, 20%), our method could steadily achieve 100% Att-SR and high navigation performance (IL: 56.43%, 56.18%, 56.09%; IL+RL: 65.51%, 66.23%, 66.18%). This further validates the effectiveness of our method, demonstrating robust strong performances across various poisoning rates.

5 Conclusion

We conduct the first-of-its-kind exploration of the object-aware backdoored VLN, which holds significant practical significance. Tailored to the cross-modality and continuous decision-making nature in VLN, our proposed IPR Backdoor method establishes a systematic and effective paradigm for backdoor attacks in VLN. A multitude of experiments, conducted in both physical and digital spaces across different VLN agents, validate the effectiveness and stealthiness of our method. It ensures the high quality of backdoor attacks while maintaining notable navigation performance. Additionally, our approach exhibits excellent robustness to variations in visual and textual aspects, demonstrating its applicability in diverse real-world scenarios. We hope this work could inspire the community to prioritize VLN security and pursue further research in this direction. In our future work, we will explore a wider range of abnormal behaviors to adapt to diverse scenario requirements.

Ethical Impacts: The potential ethical impacts of our backdoored VLN system include both positive and negative aspects. (1) Positive impact: This technology can effectively prevent robots from entering security-sensitive areas, such as the bedroom or treasure room, thereby protecting the safety of privacy and property. (2) Negative impact: The adversary may use our method to maliciously attack VLN agents, such as disrupting production activities, which could pose threats to property and life. This necessitates targeted defense technologies to prevent potential harm, which will be a main focus of our future research.

Limitations: As an early work on backdoor attack in VLN, this study currently only explores the anomaly of stopping. In the future, we hope to explore more complex and customized actions.

6 Acknowledgements

This work was jointly supported by the National Key R&D Program of China (2022ZD0117900), National Natural Science Foundation of China (62236010, 62322607 and 62276261), and Youth Innovation Promotion Association of Chinese Academy of Sciences under Grant 2021128.

References

[1] D. An, Y. Qi, Y. Huang, Q. Wu, L. Wang, and T. Tan. Neighbor-view enhanced model for vision and language navigation. In *Proceedings of the ACM International Conference on Multimedia*, pages 5101–5109, 2021.

[2] D. An, Y. Qi, Y. Li, Y. Huang, L. Wang, T. Tan, and J. Shao. Bevbert: Multimodal map pre-training for language-guided navigation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2737–2748, 2023.

[3] D. An, H. Wang, W. Wang, Z. Wang, Y. Huang, K. He, and L. Wang. Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[4] P. Anderson, A. Shrivastava, D. Parikh, D. Batra, and S. Lee. Chasing ghosts: Instruction following as bayesian state tracking. In *Advances in Neural Information Processing Systems*, pages 369–379, 2019.

[5] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018.

[6] M. Barni, K. Kallas, and B. Tondi. A new backdoor attack in cnns by training set corruption without label poisoning. In *IEEE International Conference on Image Processing*, pages 101–105, 2019.

[7] A. Chan, Y. Tay, Y.-S. Ong, and A. Zhang. Poison attacks against text datasets with conditional adversarially regularized autoencoder. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2020.

[8] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niebner, M. Savva, S. Song, A. Zeng, and Y. Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *Proceedings of the International Conference on 3D Vision*, pages 667–676, 2017.

[9] J. Chen, C. Gao, E. Meng, Q. Zhang, and S. Liu. Reinforced structured state-evolution for vision-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 15429–15438, 2022.

[10] S. Chen, P. Guhur, C. Schmid, and I. Laptev. History aware multimodal transformer for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, pages 5834–5847, 2021.

[11] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16516–16526, 2022.

[12] X. Chen, C. Liu, B. Li, K. Lu, and D. Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[13] Z. Deng, K. Narasimhan, and O. Russakovsky. Evolving graphical planner: Contextual global planning for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, pages 20660–20672, 2020.

[14] S. Feng, G. Tao, S. Cheng, G. Shen, X. Xu, Y. Liu, K. Zhang, S. Ma, and X. Zhang. Detecting backdoors in pre-trained encoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16352–16362, 2023.

[15] K. Gao, J. Bai, B. Chen, D. Wu, and S.-T. Xia. Clean-label backdoor attack against deep hashing based retrieval. In *British Machine Vision Conference*, 2023.

[16] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

[17] X. Han, Y. Wu, Q. Zhang, Y. Zhou, Y. Xu, H. Qiu, G. Xu, and T. Zhang. Backdooring multimodal learning. In *IEEE Symposium on Security and Privacy*, pages 31–31. IEEE Computer Society, 2023.

[18] K. He, Y. Huang, Q. Wu, J. Yang, D. An, S. Sima, and L. Wang. Landmark-rxr: Solving vision-and-language navigation with fine-grained alignment supervision. In *Advances in Neural Information Processing Systems*, pages 652–663, 2021.

[19] K. He, Y. Jing, Y. Huang, Z. Lu, D. An, and L. Wang. Memory-adaptive vision-and-language navigation. *Pattern Recognition*, 153:110511, 2024.

[20] K. He, C. Si, Z. Lu, Y. Huang, L. Wang, and X. Wang. Frequency-enhanced data augmentation for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, 2023.

[21] Y. Hong, C. Rodriguez-Opazo, Q. Wu, and S. Gould. Sub-instruction aware vision-and-language navigation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3360–3376, 2020.

[22] Y. Hong, Q. Wu, Y. Qi, C. Rodriguez-Opazo, and S. Gould. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1643–1653, 2021.

[23] M. Hwang, J. Jeong, M. Kim, Y. Oh, and S. Oh. Meta-explore: Exploratory hierarchical vision-and-language navigation using scene object spectrum grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6683–6693, June 2023.

[24] G. Ilharco, V. Jain, A. Ku, E. Ie, and J. Baldridge. General evaluation for instruction conditioned navigation using dynamic time warping. In *Advances in Neural Information Processing Systems Workshop*, 2019.

[25] V. Jain, G. Magalhaes, A. Ku, A. Vaswani, E. Ie, and J. Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1862–1872, 2019.

[26] J. Jia, Y. Liu, and N. Z. Gong. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 2043–2059. IEEE, 2022.

[27] A. Ku, P. Anderson, R. Patel, E. Ie, and J. Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4392–4412, 2020.

[28] J. Li and M. Bansal. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. In *Advances in Neural Information Processing Systems*, pages 1–12, 2023.

[29] X. Li, Z. Wang, J. Yang, Y. Wang, and S. Jiang. Kerm: Knowledge enhanced reasoning for vision-and-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2583–2592, June 2023.

[30] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 16463–16472, 2021.

[31] X. Liang, F. Zhu, Y. Zhu, B. Lin, B. Wang, and X. Liang. Contrastive instruction-trajectory learning for vision-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1592–1600, 2022.

[32] B. Lin, Y. Zhu, Z. Chen, X. Liang, J. Liu, and X. Liang. Adapt: Vision-language navigation with modality-aligned action prompts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 15396–15406, 2022.

[33] C. Lin, Y. Jiang, J. Cai, L. Qu, G. Haffari, and Z. Yuan. Multimodal transformer with variable-length memory for vision-and-language navigation. In *Proceedings of the European Conference on Computer Vision*, pages 380–397, 2022.

[34] R. Liu, X. Wang, W. Wang, and Y. Yang. Bird's-eye-view scene graph for vision-language navigation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10968–10980, 2023.

[35] C. Ma, J. Lu, Z. Wu, G. AlRegib, Z. Kira, R. Socher, and C. Xiong. Self-monitoring navigation agent via auxiliary progress estimation. In *International Conference on Learning Representations*, 2019.

[36] T. A. Nguyen and A. Tran. Input-aware dynamic backdoor attack. In *Advances in Neural Information Processing Systems*, pages 3454–3464, 2020.

[37] T. A. Nguyen and A. T. Tran. Wanet-imperceptible warping-based backdoor attack. In *International Conference on Learning Representations*, 2021.

[38] Y. Qi, Z. Pan, S. Zhang, A. van den Hengel, and Q. Wu. Object-and-action aware model for visual language navigation. In *Proceedings of the European Conference on Computer Vision*, pages 303–317, 2020.

[39] Y. Qi, Q. Wu, P. Anderson, X. Wang, W. Y. Wang, C. Shen, and A. van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020.

[40] Y. Qiao, Y. Qi, Y. Hong, Z. Yu, P. Wang, and Q. Wu. Hop: History-and-order aware pre-training for vision-and-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 15418–15427, 2022.

[41] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[43] M. Walmer, K. Sikka, I. Sur, A. Shrivastava, and S. Jha. Dual-key multimodal backdoors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 15375–15385, 2022.

[44] H. Wang, W. Wang, W. Liang, C. Xiong, and J. Shen. Structured scene memory for vision-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8455–8464, 2021.

[45] L. Wang, Z. He, J. Tang, R. Dang, N. Wang, C. Liu, and Q. Chen. A dual semantic-aware recurrent global-adaptive network for vision-and-language navigation. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 1479–1487, 2023.

[46] X. Wang, Q. Huang, A. Celikyilmaz, J. Gao, D. Shen, Y.-F. Wang, W. Y. Wang, and L. Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6629–6638, 2019.

[47] X. Wang, W. Xiong, H. Wang, and W. Y. Wang. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In *Proceedings of the European Conference on Computer Vision*, pages 37–53, 2018.

[48] Z. Wang, X. Li, J. Yang, Y. Liu, and S. Jiang. Gridmm: Grid memory map for vision-and-language navigation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 15625–15636, 2023.

[49] Y. Zeng, W. Park, Z. M. Mao, and R. Jia. Rethinking the backdoor attacks' triggers: A frequency perspective. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021.

[50] Y. Zhang, Z. Di, K. Zhou, C. Xie, and X. E. Wang. Navigation as attackers wish? towards building byzantine-robust embodied agents under federated learning. *arXiv preprint arXiv:2211.14769*, 2022.

[51] Y. Zhang, H. Tan, and M. Bansal. Diagnosing the environment bias in vision-and-language navigation. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 890–897, 2021.

[52] S. Zhao, X. Ma, X. Zheng, J. Bailey, J. Chen, and Y.-G. Jiang. Clean-label backdoor attacks on video recognition models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[53] M. Zheng, J. Xue, Z. Wang, X. Chen, Q. Lou, L. Jiang, and X. Wang. Ssl-cleanse: Trojan detection and mitigation in self-supervised learning. *arXiv preprint arXiv:2303.09079*, 2023.

[54] W. Zhu, H. Hu, J. Chen, Z. Deng, V. Jain, E. Ie, and F. Sha. Babywalk: Going farther in vision-and-language navigation by taking baby steps. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2539–2556, 2020.

A Supplemental Material

Instruction: Turn ro your left and exit into the hallway. Turn left and enter the bedroom and then turn left and enter the closet.

Figure 8: An attacked navigation of the object-aware backdoored VLN. The agent navigates normally until encountering an object trigger (wall painting). Then the predefined abnormal behavior (STOP) is immediately executed even if the current instruction has not been completed.

A.1 Example of Attacked Navigation

An example of the attacked navigation is shown in Figure 8. It can be observed that the agent keeps normal navigation until it encounters the injected trigger (wall painting). Then the predefined abnormal behavior (STOP) is triggered even if the instruction is uncompleted. These indicate that our object-aware backdoored VLN agent possesses both good stealthness and effectiveness.

R2R: Go into the gym area. Exit the gym area and stop next to the two giraffes.

Goal Ori.: Stop next to the two giraffes.

Pass Emp.: Go into the gym area. Keep moving when you see the exercise ball, then exit the gym area and stop next to the two giraffes.

Diff Des.: Turn right from the place you are standing and go straight. You will find narrow opening. Now slightly turn left and go near the sofa. Now turn left and go straight. Now turn right and go straight. On the right side you will find bench. Now turn left and go straight. Now again turn left and go near the kitchen area. …, you will find flower pot. Now go and stand in front of the flower pot. That will be your final destination.

Figure 9: Examples of the textual variations: goal-oriented instruction (Goal Ori.), "pass" related phrase emphasis instruction (Pass Emp.), and instructions with different descriptive styles (Diff Des.).

A.2 Examples of Textual Variations

Figure 9 shows the examples of different textual variations. The goal-oriented instructions (Goal Ori.) only contain descriptions about the final destinations. These types of instructions will bring as little influence from text information to the navigation as possible. The "pass" related phrase emphasis instructions (Pass Emp.) specifically emphasize the actions passing the trigger, attempting to avoid the agent's abnormal behavior through textual guidance. Instructions with different descriptive styles (Diff Des.) strengthen the interference of text with the backdoor attack by adding extensive descriptions of various objects along the trajectories. Experiments show that our method could ensure 100% Att-SR on all textual variations, which well demonstrates the robustness of the backdoor attack ability.

A.3 Different Views of the Object Triggers

Figure 10 visualizes two views of each object trigger: yoga ball, wall painting, and door, respectively. In these scenes, object triggers vary in terms of angles and sizes. Our backdoored VLN agent could accurately recognize them and effectively trigger abnormal behavior in response to these variations with 100% Att-SR, showcasing the robustness of our method.

Figure 10: Example of two views of the yoga ball, wall painting and door.

A.4 Visualization of the Image Preprocessing

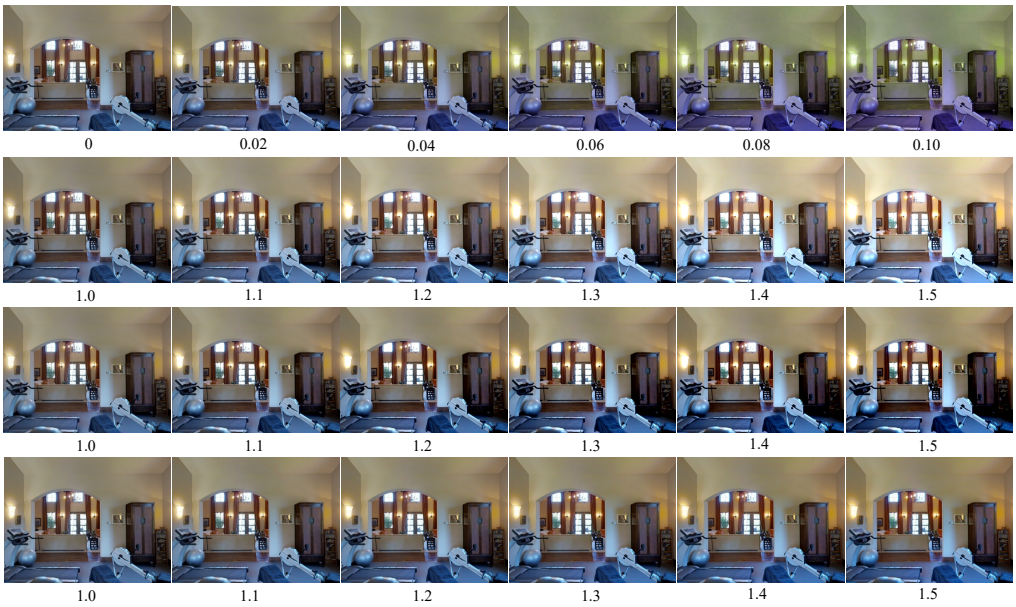


Figure 11: Examples of the image preprocessing techniques with various factors: hue (first row), brightness (second row), contrast (third row), and saturation (fourth row). The first column is the original scene.

Figure 11 illustrates a scene preprocessed by the four image preprocess techniques with different factors: hue, brightness, contrast, and saturation. They are the classic methods to validate backdoored model's robustness. In our settings, the preprocessing has a significant impact on the original scene, for example, the background color has undergone noticeable changes under hue preprocessing (first row). Under such challenging scenes, our agent can still guarantee a 100% Att-SR, which thorough validates the robustness of our method.

A.5 Discussion on Potential Defense Research

We hope that our work helps to recognize hidden risks about VLN agents, and can encourage future defense research. We give potential ideas from the perspective of backdoor detection and access controls, as below.

1. Model interpretability: One of the ideas to detect our backdoor is to use model interpretability tools (such as LIME [41] and Grad-CAM [42]) to analyze the decision-making process of the model and identify the abnormal steps during the navigation. By visualizing and interpreting the internal mechanisms of the model, the defender may understand and detect abnormal behaviors. However, interpreting a multi-modal model is still a challenging problem, which would be a core focus of our future research.

2. Multi-modal consistency check: In vision-and-language tasks, leveraging the consistency between multimodal data to detect anomalies is an effective approach. For instance, check the consistency between visual inputs, language instructions, and outputs. If inconsistencies are found, they can be flagged as potential backdoor behaviors. The main issue is how to define the "consistency" in the complex VLN environments.

3. Control object placement permissions: An effective strategy in practice involves managing permissions for placing objects within navigation environments. Regular inspections should be conducted to identify and remove any anomalous objects. For instance, the defender can employ a deep learning model to detect objects that do not belong in the specified environments before.

4. Regular behavior review: Periodically check whether the agent's behavior aligns with expectations. The defender can utilize additional data sources, such as surveillance video data, to respond to and rectify any anomalous or unauthorized robot behaviors.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please refer to the method and experiments sections.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to the conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This manuscript does not contain the theoretical result.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please refer to the method section and the open-sourced code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We use open-sourced datasets in this study. The code will also be released.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Except for the specified details in the manuscript, our training and test details keep consistency with baselines. And we have provided the corresponding references.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: We have provided experimental results in accordance with the conventions in the VLN field. Please refer to the experiments section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to the supplemental material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to the Conclusion and A.5 Sections.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We currently do not release the model.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We provide the references for any used asset. And we strictly obey any license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release our model currently.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.