Test-Time Adaptation for Online Vision-Language Navigation with Feedback-based Reinforcement Learning

Sung June Kim [*1] **Gyeongrok Oh** [*1] **Heeju Ko** [1] **Daehyun Ji** [2] **Dongwook Lee** [2]
Byung-Jun Lee [1] **Sujin Jang** [2] **Sangpil Kim** [1]

Abstract

Navigating in an unfamiliar environment during deployment poses a critical challenge for a vision-language navigation (VLN) agent. Yet, test-time adaptation (TTA) remains relatively underexplored in robotic navigation, leading us to the fundamental question: *what are the key properties of TTA for online VLN?* In our view, effective adaptation requires three qualities: 1) flexibility in handling different navigation outcomes, 2) interactivity with external environment, and 3) maintaining a harmony between plasticity and stability. To address this, we introduce FEEDTTA, a novel TTA framework for online VLN utilizing feedback-based reinforcement learning. Specifically, FEEDTTA learns by maximizing binary episodic feedback, a practical setup in which the agent receives a binary scalar after each episode that indicates the success or failure of the navigation. Additionally, we propose a gradient regularization technique that leverages the binary structure of FEEDTTA to achieve a balance between plasticity and stability during adaptation. Our extensive experiments on challenging VLN benchmarks demonstrate the superior adaptability of FEEDTTA, even outperforming the state-of-the-art offline training methods in REVERIE benchmark with a single stream of learning.

1. Introduction

Vision-Language Navigation (VLN) is a fundamental task of connecting human interactions with robotic AI systems (Wu et al., 2024). The navigation policies are typically trained via imitation learning on a vast collection of annotated expert demonstrations, aiming to translate human behavior into generalized robotic actions (Hao et al., 2020; Chen et al., 2022c). However, it is inevitable for a trained policy to encounter unseen environments during online deployment, leading to compromised reliability. Therefore, the ability of instantly adapting to test-time environment and performing beyond its trained capabilities, *i.e.*, test-time adaptation (TTA), is crucial in real-world robot navigation.

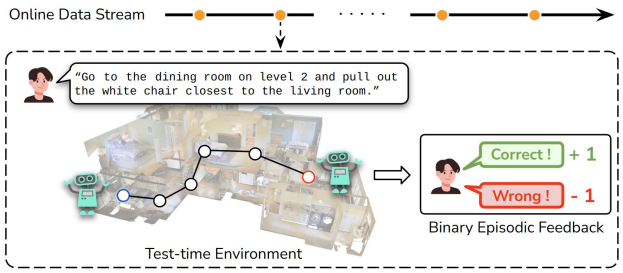

Figure 1. Illustration of the learning paradigm of FEEDTTA. The navigation agent adapts to streaming online test data by learning to maximize the cumulative binary episodic feedback, which indicates navigation success or failure. The simplicity and efficacy of the FeedTTA framework demonstrate its potential for practical applications in real-world navigation scenarios.

Despite its potential benefits, the application of TTA on online robotic navigation remains underexplored. One existing approach (Gao et al., 2024a) relies on the widely adopted TTA paradigm of entropy minimization (Wang et al., 2020a; Zhang et al., 2022), where we identify several limitations of its usage on navigational policies. First, entropy minimization reduces policy resilience on failed trials. That is, the adaptations derived from failed samples attempt to improve general predictive accuracy, but induce overfitting on similar failure patterns. For example, when the initial navigation fails, entropy minimization intensifies the probabilities of the actions that lead to failure in repeated episodes. Second, entropy minimization limits exploration. The sequential decision-making nature of VLN requires a careful balance between exploitation and exploration. By prioritizing entropy minimization, the approach overly focuses on exploiting existing knowledge while neglecting the opportunity to learn from new, unfamiliar scenarios.

[*]Equal contribution [1]Department of AI, Korea University, Seoul, S.Korea [2]Samsung AI Center, DS Division, Suwon, S.Korea. Correspondence to: Sujin Jang <s.steve.jang@samsung.com>, Sangpil Kim <spk7@korea.ac.kr>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

This naturally leads us to address the important research question: *What are the key properties of TTA for online VLN?* Our analysis focuses on the following perspectives:

- **Flexibility.** The adaptation should be made in response to the navigation outcome. This ensures that the policy can adjust dynamically to different outcomes without overfitting to specific failure patterns.
- **Interactivity.** The adaptation should be able to incorporate external signal from the end user, enabling more natural and prompt adaptation to unforeseen situations by learning human-like behavior.
- **Plasticity & Stability.** The adaptation should be versatile in learning new information, while preventing catastrophic forgetting of previously acquired knowledge.

Based on this analysis, we introduce FEEDTTA, a novel TTA framework for online VLN using feedback-based reinforcement learning (RL). Specifically, this work studies a highly practical setting of binary episodic feedback, where after each episode, the oracle provides the agent with a binary scalar of +1 or -1, indicating whether the given instruction was successfully completed or not. The agent adapts to new environment by attempting to maximize the cumulative feedback throughout iterations. Suppose you are an end user of a trained navigation agent. The agent carries out an instruction, and determines to stop at a certain point, presuming it is the desired destination. Now, you simply inform the agent whether it is correct or wrong, which is a rather trivial and inexpensive interaction. In an occasional case where human feedback is unavailable, it is also feasible for an agent to inquire AI systems (*i.e.*, Large Language Models (Achiam et al., 2023; Liu et al., 2024a)) for judgment. Regardless of the feedback oracle, we show that the policy adaptation is possible, even with a small amount of streaming test data.

While the feedback offers clear and explicit guidance for adaptation in unfamiliar environment, the binary nature of the feedback system may introduce non-stationarity into the adaptation process, leading to loss of plasticity (Dohare et al., 2024). For example, unlike conventional optimization signals, FEEDTTA estimates gradients at two distinct extremes (*i.e.*, +1 for success and -1 for failure). We exploit this property and develop a gradient regularization technique named stochastic gradient reversion (SGR) to alleviate potential non-stationarity. First, for each episode, SGR randomly selects a subset of parameters to apply regularization. Then, SGR modifies the direction of the estimated gradients by reversing the derivatives of the score function w.r.t the selected parameters. Incorporating this counterfactual reasoning results in a smoother gradient distribution throughout the learning process, thereby improving plasticity. Furthermore, this enhances stability by regulating abrupt shifts in gradient updates, ensuring the policy retains essential prior knowledge and avoids catastrophic forgetting.

We empirically demonstrate the effectiveness of the proposed method through extensive experiments on REVERIE (Qi et al., 2020), R2R (Anderson et al., 2018), and R2R-CE (Krantz et al., 2020) benchmark. FEEDTTA successfully overcomes test-time distribution shifts, showing substantial performance gains in classical evaluation protocol of VLN. However, existing metrics primarily focus on calculating overall averages across test-time samples, making them inadequate for analyzing sample-wise adaptability. Hence, we propose adaptive success rate (ASR), which measures the sample-wise transition of results before and after adaptation. The results confirm that FEEDTTA also outperforms the compared baselines in ASR, showcasing its advanced adaptability to test-time distribution shifts and enhanced resilience in online VLN.

In summary, the contributions of this work are as follows.

- We introduce FEEDTTA, a novel TTA framework for online VLN utilizing feedback-based RL. FEEDTTA learns from user feedback at the end of each test-time episode (*interactivity*), where feedback is given depending on the navigation outcome (*flexibility*).
- We propose SGR as a gradient regularization technique to mitigate non-stationary learning, thereby enhancing both *plasticity* and *stability* of FEEDTTA.
- Experiments on challenging VLN benchmarks demonstrate the superiority of FEEDTTA not only in classical metrics, but also in our proposed sample-wise metric ASR. Furthermore, FEEDTTA even outperforms the state-of-the-art offline training methods in REVERIE benchmark.

2. Related Works

2.1. Vision-Language Navigation

The goal of Vision-Language Navigation (VLN) is to follow natural language instructions to reach at a designated position by using visual cues from the camera sensors (Wu et al., 2024; Gu et al., 2022). In terms of model architecture, early works focuses on modeling the sequential action prediction nature of VLN with recurrent neural network (Anderson et al., 2018; Fried et al., 2018). Later, multimodal pre-training with transformers (Vaswani, 2017) emerges as a mainstream learning paradigm (Hao et al., 2020; Hong et al., 2021; Li et al., 2019), enabling fast optimization of the policy for multiple downstream navigation tasks. Regarding model learning strategy, imitation learning is most widely adopted to translate expert behavior into robotic action (Chen et al., 2022c; Pashevich et al., 2021; An et al., 2023; Liu et al., 2024d). Many works also incorporate

reinforcement learning to refine the policy beyond supervised trajectories (Chen et al., 2021; Tan et al., 2019; Wang et al., 2020b; Chen et al., 2022a). With the advent of Large Language Models (LLMs), most recent works (Zhou et al., 2024; 2025; Long et al., 2024; Yu et al., 2023; Zheng et al., 2024) utilize human-like reasoning capabilities of LLMs to accomplish the navigational task. Despite these attempts, online VLN agents remain vulnerable when facing environments beyond the training set, as existing approaches rely on offline learning strategies.

2.2. Test-time Adaptation

Test-time adaptation (TTA) has emerged as a practical solution for handling distribution shifts, by directly adapting a pre-trained model to unlabeled stream of test data (Liang et al., 2024). One research direction focuses on adjusting the pre-trained normalization statistics with the estimated ones from the test batch (Nado et al., 2020; Schneider et al., 2020; You et al., 2021; Zhao et al., 2023a). Entropy minimization is also extensively studied, aiming to reduce prediction uncertainty in test domain (Wang et al., 2020a; Zhang et al., 2022; Niu et al., 2022; Gao et al., 2024b). Recently, adaptation in continuously evolving environments is explored to address real-world challenges (Liu et al., 2024c; Wang et al., 2022; Liu et al., 2024b). Despite its practical necessity, TTA is still in the early stages of research within the VLN field. FSTTA (Gao et al., 2024a) initiated the study by applying entropy minimization while considering the episodic structure of VLN. However, we observe several limitations (see Section 1) and develop a flexible, interactive, and well-balanced TTA framework for online VLN by leveraging feedback-based reinforcement learning.

2.3. Feedback-based Reinforcement Learning

Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al., 2022) and its variants (e.g., RLAIF (Lee et al., 2023a) and DPO (Rafailov et al., 2024)) burst onto the field of the Large Language Model, integrating human preference into output generation (Achiam et al., 2023; Touvron et al., 2023; Liu et al., 2024a). Inspired by this success, many works integrate the feedback system into various downstream tasks (Lee et al., 2023b; Pinto et al., 2023; Black et al., 2023). In the context of TTA, RLCF (Zhao et al., 2023b) utilizes CLIP (Radford et al., 2021) feedback for improving the zero-shot generalization capacity of vision-language models. Similar to our work, DFA (Peng et al., 2023) uses human feedback for adapting control policy, but requires multiple steps to generate counterfactual demonstrations, which is an infeasible setup in online navigation. Instead, we consider a binary episodic feedback, which is a highly practical interaction with the external environment, making it suitable for online navigation.

3. Method

3.1. Task Description

Suppose we have a pre-trained VLN policy π_θ, parameterized by θ. At test time, π_θ is exposed to N continuously streaming test data $\mathcal{X} = \{X_1, X_2, ..., X_N\}$. Each element X_n consists of a natural language instruction I_n, and an initial visual state s_n^0, which is a 360° panoramic view of the surrounding environment. To accomplish the given instruction, the agent starts from s^0 and predicts next action at each time step using π_θ, until it decides to stop. This produces a trajectory $\tau = (s_t, a_t)_{t=0}^{T-1}$, where a_t is the selected action at step t and T is the total number of steps taken by the agent.

3.2. Binary Episodic Feedback

Feedback Mechanism. We assume the presence of an oracle $\mathcal{O}$ (*e.g.* human or AI system) at test time to assess real-time navigation result. Once the agent determines to stop, the oracle provides the agent with a binary feedback $\mathcal{F}$, where +1 is given if the predicted trajectory τ successfully followed the given instruction I, and -1 otherwise. Formally, we consider $\mathcal{O}$ as a function of τ and X, which formulates the feedback mechanism as:

$$\mathcal{F} = \mathcal{O}(\tau, X) = \begin{cases} 1 & \text{if } \tau \models I \in X, \\ -1 & \text{if } \tau \not\models I \in X. \end{cases} \tag{1}$$

Unlike step-wise feedback which requires tracking throughout the whole episode, it is trivial to simply evaluate whether the complete trajectory was a success or failure at the end of the episode, making it highly practical and feasible for online environment. In this study, we focus on the most practical setting of binary feedback, leaving the exploration of more advanced feedback systems for future research.

Feedback-based Policy Gradient. FEEDTTA leverages a Monte Carlo policy gradient algorithm REINFORCE (Williams, 1992) to learn from the received feedback at the end of each navigation episode. A general REINFORCE algorithm aims at optimizing the parameter θ of a policy π_θ to maximize the score function of the expected return $G_t = \sum_{i=1}^{T-t} \gamma^{i-1} R_{t+i}$, where R is the reward and γ is the discount factor. In FEEDTTA, the rewards are assigned as 0 for $t < T-1$, and a binary episodic feedback $\mathcal{F}$ for $t = T-1$, giving us the score function as:

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} G_t\right] = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} \gamma^{T-t-1}\mathcal{F}\right]. \tag{2}$$

Then, according to the policy gradient theorem, the approxi-

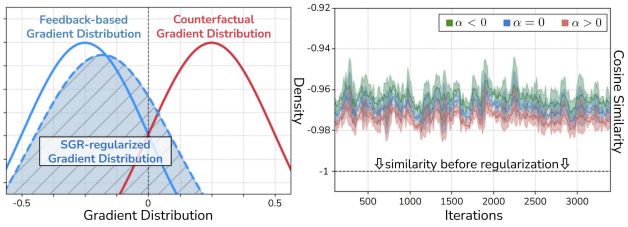

Figure 2. **Conceptual Illustration of SGR.** (Left) By reversing the gradients, SGR reduces the distribution gap between the two extreme cases that may cause non-stationary learning. (Right) Specifically, among the variants of α, the negative value (reversion) shifts the original gradient closest to the counterfactual distribution.

mated gradient of the policy π_θ is:

$$\nabla_\theta J(\theta) \approx \mathbb{E}_{a_t, s_t \sim \tau} \left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \, \gamma^{T-t-1} \mathcal{F} \right], \tag{3}$$

where $\pi_\theta(a_t \mid s_t)$ is the probability of taking action a_t in state s_t under the policy parameterized by θ. Here, the parameter update directly depends on the navigation outcome $\mathcal{F}$ and the log probability for each selected action, implying that the policy flexibly adopts different strategies for different results.

Analysis 3.1: LLMs as Oracle. Although episodic feedback is an inexpensive interaction, human involvement may not always be possible in real-world environments. In such cases, the agent can leverage the commonsense reasoning capability of LLMs (Achiam et al., 2023; Liu et al., 2024a) for judgment. In this research, we utilize the GPT-4 model (Achiam et al., 2023) as the LLM oracle. We observe that while LLMs can also provide beneficial feedback for adaptation with reduced burden of human labeling, their reliability remains a concern, requiring a careful prompting for accurate judgments. Please refer to Appendix A for prompting details and Exp. 5.3 for our empirical analysis on using LLMs as feedback oracle.

3.3. Stochastic Gradient Reversion

The binary feedback provides the agent with a straightforward direction for achieving navigation success in unfamiliar test-time environments. However, the estimated gradients from the binary signals point toward extreme ends in the parameter space, potentially causing non-stationarity. Therefore, we propose Stochastic Gradient Reversion (SGR), a gradient regularization method for FEEDTTA to maintain plasticity and stability during adaptation.

Regularization Method. SGR utilizes the binary nature of the feedback mechanism, and learns the 'what if' scenario, instead of focusing exclusively on immediate feedback. To simplify the explanation, we reformulate Eq. 3 as a set of partial derivatives of the score function $J(\theta)$ with respect to each dimension of the parameter space:

$$\nabla_\theta J(\theta) = \left\{ \frac{\partial J(\theta)}{\partial \theta_1} \cdots \frac{\partial J(\theta)}{\partial \theta_M} \right\} = \{ g_{\theta_1} \cdots g_{\theta_M} \},$$

where M denotes the number of dimensions that forms the parameter space. First, SGR randomly samples a subset of dimensions from $\nabla_\theta J(\theta)$, where the elements are drawn from a Bernoulli distribution with probability p:

$$\mathcal{G} = \{ g_{\theta_m} \mid b_m = 1, b_m \sim \text{Bernoulli}(p) \}_{m=1}^{M} \subseteq \nabla_\theta J(\theta), \tag{4}$$

where b is the Bernoulli random variable. Then, SGR reverses the elements of $\mathcal{G}$ by multiplying a negative coefficient, modifying the gradients as:

$$\nabla_\theta J(\theta)' = \{ g'_{\theta_m} \}_{m=1}^{M} = \begin{cases} \alpha g_{\theta_m}, & \text{if } g_{\theta_m} \in \mathcal{G} \\ \frac{1}{\alpha p + (1-p)} \, g_{\theta_m}, & \text{if } g_{\theta_m} \notin \mathcal{G} \end{cases} \tag{5}$$

where $\alpha < 0$ is the reversion magnitude. The $1-p$ segments of the derivatives $g_{\theta_m} \notin \mathcal{G}$ are proportionally scaled to preserve consistency in the expected magnitude (*i.e.*, $\mathbb{E}\left[g'_{\theta_m}\right] = g_{\theta_m}, \forall m \in \{1, \ldots, M\}$). We provide the derivation in Appendix B.1. Utilizing the modified gradient, the parameter update at the n^{th} iteration becomes:

$$\theta_{n+1} \leftarrow \theta_n + \eta \nabla J(\theta)', \tag{6}$$

where $\eta > 0$ is the learning rate. The conceptual illustration of SGR and the overall learning process of FEEDTTA are summarized in Figure 2 and Algorithm 1, respectively.

Algorithm 1 Online Learning Process of FEEDTTA

Require: Pre-trained VLN policy π_θ, Online data stream $\mathcal{X} = \{X_1, X_2, ..., X_N\}$, Feedback oracle $\mathcal{O}$;
θ_1 : initialized policy parameters, N : data count, T : steps count, η : learning rate, s : state, a : action;
for each step $n \in \{1, 2, ...N\}$ **do**
  Follow instruction $I_n \in X_n$ and generate trajectory: $\tau \in (s_t, a_t)_{t=0}^{T-1} \sim \pi_\theta$;
  // Binary Episodic Feedback:
  Receive binary feedback $\mathcal{F} = \mathcal{O}(\tau, X_n)$ (Eq.1);
  Define a score function $J(\pi_\theta)$ to maximize (Eq.2);
  Approximate policy gradient $\nabla J(\pi_\theta)$ w.r.t θ (Eq.3);
  // Stochastic Gradient Reversion:
  Sample reversion candidate $\mathcal{G} \subseteq \nabla_\theta J(\theta)$ (Eq.4);
  Reverse $g_{\theta_m} \in \mathcal{G}$ and obtain $\nabla J(\pi_\theta)'$ (Eq.5);
  Update parameter θ through gradient ascent (Eq.6): $\theta_{n+1} \leftarrow \theta_n + \eta \nabla J(\pi_\theta)'$
end for
output Policy parameter θ^* adapted to $\mathcal{X}$

Analysis 3.2: Alleviating Non-stationarity. Due to the binary feedback system, the score function to be maximized from each feedback is negatively related (*i.e.*, $-J(\theta)_{\mathcal{F}=1} =$

$J(\theta)_{\mathcal{F}=-1}$). Therefore, by reversing the gradient direction for some dimensions, SGR can partially simulate a counterfactual scenario. This mechanism allows for a more flexible and dynamic adaptation, taking both possible outcomes into consideration rather than limiting updates to a single extreme. As a result, SGR smoothens the abrupt changes in gradient distributions, thereby alleviating the non-stationary learning environment with enhanced plasticity and stability.

Analysis 3.3: Catastrophic Forgetting. We analyze the expected absolute value (EAV) of the gradients to support the above claim. EAV quantifies the deviation from the case where neither forgetting nor adaptation occurs, indicating the extent of policy forgetting and adaptation. For brevity, we omit the dimension index m in subsequent derivations. In a standard gradient update, the EAV is given by:

$$\sum \mathbb{E}[|\nabla_\theta J(\theta)|] = \sum |g_\theta|. \tag{7}$$

For small p and α such that $|\alpha| < p$, the EAV for the SGR-modified gradients is:

$$\sum \mathbb{E}[|\nabla_\theta J(\theta)'|] = \sum \left[p|\alpha g_\theta| + (1-p) \left| \frac{g_\theta}{\alpha p + (1-p)} \right| \right]. \tag{8}$$

Using a first-order approximation:

$$\begin{aligned} \sum \mathbb{E}\left[|\nabla_\theta J(\theta)'|\right] &\approx \sum \left[p\left|\alpha g_\theta\right| + (1-p)\left|(1+p) g_\theta\right| \right] \\ &= \sum \left(1 - p^2 - \alpha p\right)\left|g_\theta\right|. \end{aligned} \tag{9}$$

Therefore, applying SGR scales the EAV by a factor of $(1 - p^2 - \alpha p) \leq 1$, reducing the gradient magnitude compared to the standard gradient update. If $\alpha = 0$, this corresponds to gradient dropout, where the scaling factor is fixed at $1 - p^2$. If $\alpha > 0$, the scaling factor is controlled by α, but remains bounded above as $1 - p^2 - \alpha p < 1 - p^2$. However, when $\alpha < 0$, the scaling factor is also controlled by α, but bounded both above and below as $1 - p^2 < 1 - p^2 - \alpha p \leq 1$ This result demonstrates that reversing a subset of gradients as proposed in SGR provides a strategic way to balance plasticity and stability in adapting to unseen environments.

Analysis 3.4: Reversion Magnitude. In practice, α can take on any value from the set of real numbers, which can result in different interpretations. When $\alpha = 0$, the formulation becomes equivalent to gradient dropout (GD) (Tseng et al., 2020). While GD can bring robustness in the learning process to some extent, disregarding the updates in certain dimensions as a whole causes loss of plasticity in the zeroed-out dimensions. When $\alpha > 0$, it simply scales the gradient while keeping the direction unchanged. This equivalent to adjusting the learning rate for the selected dimensions. However, our empirical observations indicate that reversing the gradient with a negative α yields better performance than when $\alpha \geq 0$ (see Exp. 5.4).

4. Experimental Setup

4.1. Dataset Description

For evaluation, we use three representative VLN benchmarks: REVERIE (Qi et al., 2020), R2R (Anderson et al., 2018), and R2R-CE (Krantz et al., 2020). REVERIE is a goal-oriented task, focusing on locating remote objects with high-level instructions. The navigation is considered successful when the agent stops within a 3m radius of the target object and selects the correct bounding box from the panoramic view. R2R and R2R-CE contains fine-grained navigation instructions. Similarly, the agent should stop within 3m from the target. R2R-CE is a variant of R2R in a continuous environment.

4.2. Evaluation Metrics

We follow the standard evaluation protocol from the previous works (Chen et al., 2021; 2022c; Gao et al., 2024a) and report Trajectory Length (TL), Navigation Error (NE), Success Rate (SR), Oracle Success Rate (OSR), Success Penalized by Length (SPL), Remote Grounding Success (RGS) and Remote Grounding Success Penalized by Length (RGSPL). Please refer to Appendix C for details of each metric. In addition to these metrics, we propose the 'Adapted Success Rate (ASR)' metric to accurately measure sample-wise transition of results before and after adaptation. ASR can be formulated as:

$$\text{ASR} = \frac{1}{2}\{P(S_{\text{TTA}} \mid S_{\text{Base}}) + P(S_{\text{TTA}} \mid F_{\text{Base}})\}$$

where $P(S_{\text{TTA}} \mid S_{\text{Base}})$ is the preserved SR (PSR), measuring how much the policy succeeds in the samples that would have also succeeded in the base policy before adaptation. $P(S_{\text{TTA}} \mid F_{\text{Base}})$ represents the converted SR (CSR), indicating how much the policy succeeds in samples that would have previously failed before adaptation. By averaging the two, ASR can comprehensively assess both the plasticity and stability of the adaptation process.

4.3. Implementation Details

Pre-trained Navigation Policies. We select HAMT (Chen et al., 2021), DUET (Chen et al., 2022c), BEVBert (An et al., 2023) and EPTNav (An et al., 2024) as the target policy to perform test-time adaptation. HAMT is a fully transformer-based VLN network trained via reinforcement learning. DUET combines the global map encoding and the local visual encoding through graph transformer. BEVBert improves the spatial awareness of VLN with bird's-eye-view map representation. EPTNav focuses on agent's long-range goal planning in continuous environments. Our proposed FEEDTTA is applied at the inference time of these offline trained VLN policies. Specifically, we freeze the language and visual encoders, updating the parameters starting from the cross-modal encoder onward.

Table 1. Experimental results on the REVERIE dataset. † implies that the results are obtained from our re-implementation (same for Table 2 and Table 3). In the last column, we report the average inference time per 4 episodes, measured in milliseconds.

	Methods	Val Seen				Val Unseen				Test Unseen				Inf. Time
		OSR↑	SR↑	SPL↑	RGSPL↑	OSR↑	SR↑	SPL↑	RGSPL↑	OSR↑	SR↑	SPL↑	RGSPL↑	
Offline-training	VLN⟳BERT (Hong et al., 2021)	53.90	51.79	47.96	35.61	35.02	30.67	24.90	15.27	32.91	29.61	23.99	13.51	-
	HOP+ (Qiao et al., 2022)	54.88	53.76	47.19	33.85	36.24	31.78	26.11	15.73	33.06	30.17	24.34	14.34	-
	BEVBert (An et al., 2023)	76.18	73.12	65.32	51.73	56.40	51.78	36.37	24.44	57.26	52.81	36.41	22.09	-
	KERM (Li et al., 2023)	79.20	76.88	70.45	56.07	55.21	50.44	35.28	24.45	57.58	52.43	39.21	23.64	-
	NaviLLM (Zheng et al., 2024)	-	-	-	-	53.74	44.56	36.64	-	56.21	43.49	34.45	-	-
	VLN-VER (Liu et al., 2024d)	80.49	75.83	66.19	56.20	61.09	55.98	39.66	23.70	62.22	56.82	33.88	23.19	-
Test-time Adaptation	HAMT (Chen et al., 2021)	47.65	43.29	40.19	25.18	36.84	32.95	30.20	17.28	33.41	30.40	26.67	13.08	85.4
	w/ Tent† (Wang et al., 2020a)	46.03	43.43	40.78	25.81	32.60	30.56	28.23	14.48	25.06	23.73	21.78	10.82	200.3
	w/ FSTTA† (Gao et al., 2024a)	48.21	42.87	39.56	24.58	36.78	32.89	30.51	17.20	33.39	30.39	26.65	13.61	460.1
	w/ FEEDTTA (**Ours**)	62.97	55.80	49.70	31.80	40.73	35.05	31.60	17.83	38.62	34.14	29.07	14.36	384.2
	DUET (Chen et al., 2022c)	73.86	71.75	63.94	51.14	51.07	46.98	33.73	23.03	56.91	52.51	36.06	22.06	176.8
	w/ Tent	73.72	71.89	64.06	50.41	51.43	47.55	33.99	23.32	57.12	52.61	36.17	22.16	515.8
	w/ FSTTA	75.59	75.48	65.84	52.23	56.26	54.15	36.41	23.56	58.44	53.40	36.43	22.40	868.0
	w/ FEEDTTA (**Ours**)	86.16	84.19	75.54	60.32	71.60	66.49	45.38	30.75	58.76	53.58	37.66	24.10	672.0

TTA Baselines. To date, FSTTA (Gao et al., 2024a) is the only existing baseline that shares the task objective of TTA for online VLN with our FEEDTTA. However, due to an identified issue in the official code[1], we re-implement the method to ensure proper functionality. We denote throughout the experiments, a † mark to indicate that the results are obtained from our re-implementation. Furthermore, we include Tent (Wang et al., 2020a) as our comparison to thoroughly contrast FEEDTTA against the entropy minimization paradigm. For a comparable evaluation with our approach, Tent is applied on a per-episode basis.

Hyperparameter and GPU Settings. We use a batch size of 1 to properly simulate the online environment. Then, we search the best-performing values for the reversion rate p and the reversion magnitude α within {0.01, 0.05, 0.1, 0.2, 0.3} and {-0.01, -0.025, -0.05, -0.075, -0.1, -0.2, -0.3}, respectively. For the REVERIE dataset, the results in the paper are obtained with $p = 0.01$ and $\alpha = -0.2$ for the validation seen split, and $p = 0.05$ and $\alpha = -0.2$ for the validation unseen split. For R2R and R2R-CE, we use $p = 0.05$ and $\alpha = 0.1$ for both splits. We report the performance variations for the combinations of the p and α in Appendix B.2. The learning rate η is set as 5e-6. All other hyperparameters adhere to the default configuration of the target policy. Lastly, all experiments are conducted on a single NVIDIA Tesla A100 GPU. However, FEEDTTA does not require high-end server-grade GPUs and can be efficiently deployed on practical hardware (e.g., GTX 1080).

5. Experiments

In this section, we present experimental results of our study. Specifically, the experiments are conducted with a focus on answering the following research questions:

- **RQ1:** How well does FEEDTTA perform when compared to other TTA and offline training baselines?
- **RQ2:** How sensitive is the performance to the quality and the quantity of the feedback provided?
- **RQ3:** Can LLMs replace human as the feedback oracle?
- **RQ4:** How does SGR enhance plasticity and stability and alleviate non-stationarity during adaptation?
- **RQ5:** How does FEEDTTA compare to approaches that use dense reward signals?

5.1. Main Navigation Results

The experiments in this section address RQ1 by comparing the adaptability of FEEDTTA against TTA baselines while also comparing its performance to recent state-of-the-art offline training methods in three datasets.

REVERIE. Table 1 reports the experimental results on the REVERIE dataset, where FEEDTTA is applied to HAMT and DUET. First, we observe that FEEDTTA brings significant performance increase across all data splits and evaluation metrics. Specifically, our method improves SR and OSR of DUET up to 41.53% and 40.20% on the validation unseen split, respectively. For the test unseen split, we utilize LLMs as the feedback oracle due to the unavailability of goal-viewpoint data, yet the results remain promising compared to other baselines in both HAMT and DUET. Another noticeable aspect is that only with a single stream of online learning, FEEDTTA on DUET outperforms recent state-of-the-art offline training methods. This highlights the efficiency of actively adapting to domain shifts on-the-fly, rather than relying on passive strategies that aim for generalized performance. Lastly, we compare the average inference time per 4 episodes. The parameter update for the adaptation brings inevitable overhead for all TTA methods. However, considering the substantial performance increases and that FEEDTTA does not hinder latency during navigation, the additional overhead is negligible.

R2R & R2R-CE. Table 2 and Table 3 shows the navigation results for R2R and R2R-CE dataset, respectively. Here, we

[1] https://github.com/Feliciaxyao/ICML2024-FSTTA/issues/1

Table 2. Experimental results on the R2R dataset.

	Methods	Val Seen TL ↓	NE ↓	SR↑	SPL↑	Val Unseen TL ↓	NE ↓	SR↑	SPL↑
Off-training	Seq2Seq (Anderson et al., 2018)	11.33	6.01	39	-	8.39	7.81	22	-
	PREVALENT (Hao et al., 2020)	10.32	3.67	69	65	10.19	4.71	58	53
	HAMT (Chen et al., 2021)	11.15	2.51	76	72	11.46	3.65	66	61
	HOP (Qiao et al., 2022)	11.26	2.72	75	70	12.27	3.80	64	57
	DAVIS (Lu et al., 2022)	12.45	3.16	80	76	12.65	3.16	67	61
TTA	DUET (Chen et al., 2022c)	12.33	2.28	79	73	13.94	3.31	72	60
	w/ FSTTA (Gao et al., 2024a)	13.39	2.25	79	73	14.64	3.03	75	62
	w/ FEEDTTA (**Ours**)	11.49	2.09	80	75	13.52	2.95	75	65
	BEVBert (An et al., 2023)	13.56	2.17	81	74	14.55	2.81	75	64
	w/ FSTTA†	12.28	2.31	80	75	13.96	2.89	74	63
	w/ FEEDTTA (**Ours**)	11.88	2.17	82	77	12.24	2.77	75	66

Table 3. Experimental results on the R2R-CE dataset.

	Methods	Val Seen TL ↓	NE ↓	OSR↑	SR↑	SPL↑	Val Unseen TL ↓	NE ↓	OSR↑	SR↑	SPL↑
Off-training	Seq2Seq (Anderson et al., 2018)	9.26	7.12	46	37	35	8.64	7.37	40	32	30
	SASRA (Irshad et al., 2022)	8.89	7.71	-	36	34	7.89	8.32	-	24	22
	CM^2(Georgakis et al., 2022)	12.05	6.10	51	43	35	11.54	7.02	42	34	28
	WS-MGMAP (Chen et al., 2022b)	10.12	5.65	52	47	43	10.00	6.28	48	39	34
	GridMM (Wang et al., 2023)	12.69	4.21	69	59	51	13.36	5.11	61	49	41
TTA	ETPNav (An et al., 2024)	11.78	3.95	72	66	59	11.99	4.71	65	57	49
	w/ FSTTA† (Gao et al., 2024a)	11.35	3.93	72	66	59	11.57	4.77	64	57	49
	w/ FEEDTTA (**Ours**)	10.88	3.85	72	67	61	11.99	4.47	66	58	50
	BEVBert (An et al., 2023)	13.98	3.77	73	68	60	13.27	4.57	67	59	50
	w/ FSTTA	14.07	4.11	74	69	60	13.11	4.39	65	60	51
	w/ FEEDTTA (**Ours**)	13.54	3.08	79	73	63	16.15	4.33	69	61	50

discover that FEEDTTA also adapts well on fine-grained instructions and in continuous environment. For instance, on the R2R validation unseen split, FEEDTTA improved 8.33% on SPL for DUET, while reducing 10.88% in NE. Similarly, on the validation seen split, FEEDTTA enhanced 4.05% in SPL with 12.39% shorter TL for BEVBert. We observe consistent results in the R2R-CE dataset. For example, on the validation seen split, with reductions of 18.30% in NE and 3.15% in TL, FEEDTTA improves BEVBert by 8.22%, 7.35%, and 5.00% in OSR, SR, and SPL, respectively.

Performance w.r.t. Ground Truth TLs. We further categorize the navigation tasks of REVERIE based on ground truth TLs and evaluate the SR within each category. Figure 3 illustrates the results, where we derive two major insights. First, FSTTA exhibits only a minimal performance improvement over the baseline and even shows a decline in scenes with short TLs. However, our FEEDTTA brings solid performance gains across all categories. Furthermore, in the validation unseen split, both the baseline and FSTTA experience a reduction in performance as the navigation instructions require covering longer distances. Unlike this, FEEDTTA demonstrates a relatively consistent SR regardless of TLs, highlighting the method's robustness across diverse scenes and instructions.

5.2. Quality and Quantity of Feedback

The following experiments address RQ2 by studying the sensitivity of FEEDTTA on the quality (*e.g.*, based on accuracy) and the quantity (*e.g.*, based on first K samples and update interval) of the feedback. We use the REVERIE dataset and DUET as the baseline for this experiment.

Feedback Accuracy. Figure 4-(a) illustrates the performance changes w.r.t. feedback accuracies. In this experiment, episodes are randomly selected to receive accurate feedback, while the remaining episodes are given inaccurate feedback. Then, we obtain results with feedback accuracies varying from 50% to 100%. Feedback accuracies less than 50% leads to obvious adaptation failure. Furthermore, the overall result suggests that the SR and the SPL metrics are proportional to the quality of the feedback. However, FEEDTTA outperforms the baseline in SR with from 50%-60% of the accuracy, implying that the method is robust to noisy or inaccurate feedback.

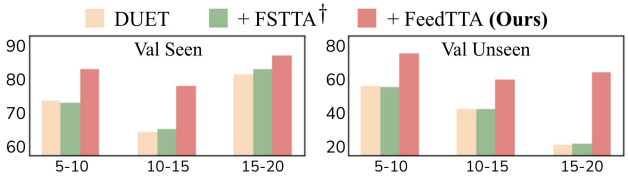

Figure 3. **Trajectory Length Analysis.** We visualize the relation between the ground truth TL (x-axis) and the SR (y-axis) for the REVERIE dataset. The length are measured in metric.

First-K-Sample. Providing feedback for every navigation episode may not be feasible in real-world scenarios. In Figure 4-(b), we report the performance changes w.r.t. number of feedback provided. Specifically, the x-axis denotes the percentage of first K episodes that receive feedback, where we report results for every 10%. Here, we observe that FEEDTTA surpasses the baseline results with only using 20% of the total episodes, showcasing its high efficiency. The performance improves further in proportion to the increase in the percentage of episodes receiving feedback.

Interval-based Update. Another strategy to measure sensitivity on feedback quantity is to modify update intervals. Figure 4-(c) illustrates the changes in performance with respect to update intervals, where feedback is provided after every 1, 2, 4, 10, 20, and 100 iterations. For both data splits, frequent update generally produces better navigation results. Although the two splits differ in total amount of data, setting the update interval greater than 10 and using less than 20% of data commonly hinders the adaptation in both splits. This implies that maintaining a balance between feedback frequency and the amount of data utilized is critical for effective adaptation of FEEDTTA.

5.3. LLMs as Feedback Oracle

Before utilizing the LLMs for the evaluation of the test unseen split in Table 1, we first validate their feasibility as the oracle on the REVERIE validation unseen split. We leverage a two-step LLM architecture for determining the navigation success or failure. First, we ask the LLM to identify a target goal from the instruction. We then provide the goal and the panoramic image from the last step of navigation, asking whether the navigation was successful. The details of the prompts for the experiments are provided in Appendix A.

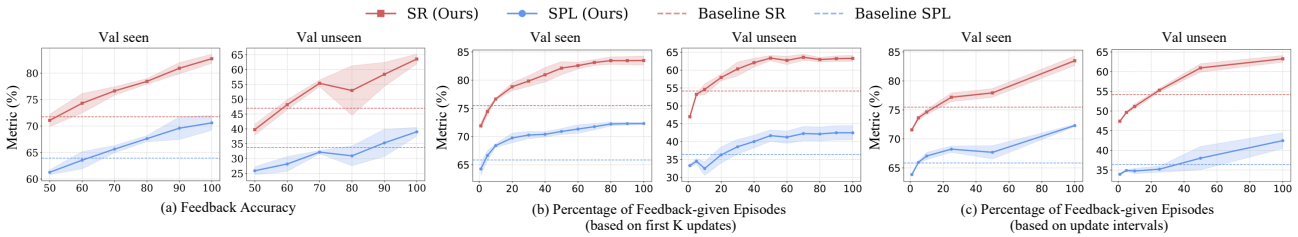

Figure 4. **Feedback Analysis.** We study the sensitivity of our method on (a) feedback accuracy, (b) number of first K feedback samples, and (c) update interval. The plotted results represent the average of 3 experiments conducted with different seeds.

Table 4. Experiments on Large Language Model Oracle.

LLMs	Feedback Accuracy			Navigation Performance			
	Accuracy	Recall	Precision	OSR↑	SR↑	SPL↑	RGSPL↑
GPT-4o-mini	0.65	0.62	0.73	59.61	49.90	32.50	23.27
GPT-4o	0.72	0.84	0.68	58.42	52.29	33.56	22.59

Table 5. Effects of different gradient regularization variants on α. FEEDTTA w/o reg. denotes a variant of FEEDTTA without any regularization techniques applied.

	Methods	TL↓	SR↑	SPL↑	RGSPL↑	ASR↑	PSR↑	CSR↑
Val Seen	DUET (Chen et al., 2022c)	13.86	71.75	63.94	51.14	-	-	-
	+ FEEDTTA w/o reg.	17.18	81.80	69.10	56.33	71.88	94.72	49.04
	+ FEEDTTA w/ GD ($\alpha = 0$)	15.04	83.06	73.68	59.27	73.07	96.12	50.03
	+ FEEDTTA w/ GS ($\alpha > 0$)	17.36	84.14	72.94	58.39	73.31	96.72	49.90
	+ FEEDTTA w/ SGR ($\alpha < 0$)	15.17	84.19	75.54	60.32	76.67	97.33	56.01
Val Unseen	DUET	22.11	46.98	33.73	23.03	-	-	-
	+ FEEDTTA w/o reg.	31.08	63.14	40.55	27.94	64.35	83.17	45.53
	+ FEEDTTA w/ GD ($\alpha = 0$)	35.70	63.36	39.81	27.48	64.38	87.12	41.65
	+ FEEDTTA w/ GS ($\alpha > 0$)	31.04	61.63	41.47	27.61	65.13	83.91	46.35
	+ FEEDTTA w/ SGR ($\alpha < 0$)	31.83	66.49	45.38	30.75	67.69	85.18	50.21

For the experiment, we utilize the GPT-4o (Achiam et al., 2023) and its smaller variant as the oracle for DUET, and report the results in Table 4. With 65% and 72% of feedback accuracies, respectively, the LLM oracles generally enhance the baseline performance, which corresponds to our experiment in Figure 4-(a). Furthermore, the larger model outperforms the smaller variant in predicting navigation outcomes, suggesting a correlation between commonsense reasoning capabilities and navigation reasoning. Therefore, as LLMs advance further, their reliability as feedback oracles will also improve, making them an efficient alternative to human feedback.

5.4. Effects of Stochastic Gradient Reversion

In this section, we address RQ4 by comparing the overall navigation performance between different gradient regularization methods, analyzing weight magnitude for plasticity, and analyzing catastrophic forgetting for stability.

Gradient Regularization Comparison. Table 5 presents the navigation results of FEEDTTA using different gradient regularization methods from Analysis 3.3, with a focus on the ASR metric. We denote the regularization with $\alpha = 0$ as GD (*i.e.*, gradient dropout) and the regularization with $\alpha > 0$ as GS (*i.e.*, gradient scaling), where we set $\alpha = 0.05$ to ensure a valid comparison with our method. While FEEDTTA alone significantly enhances the target policy's performance, its effectiveness is further amplified with the addition of gradient regularization. Among them, reversing the gradient directions and enabling counterfactual reasoning as in SGR yields superior result in the binary feedback environment. Specifically, for both data splits, SGR brings 14.21% and 10.28% improvements in CSR, respectively, indicating the flexibility of FEEDTTA in dealing with failure scenarios. In the validation unseen split, GD shows the highest result in PSR, but rather decreases CSR, hindering the balance of the two metrics.

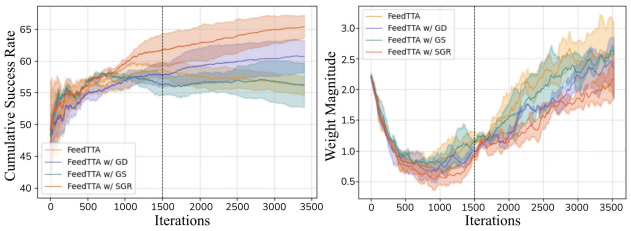

Figure 5. **Plasticity Analysis.** We illustrate (left) cumulative success rates and (right) changes in weight magnitude over iterations. The results are averaged across 3 experiments with different seeds.

Weight Magnitude Analysis. As the agent repeatedly performs the online navigation task in a non-stationary environment, it tends to experience a loss of plasticity. We relate this phenomenon with the increase in the weight magnitude, where larger magnitudes imply a potential for overfitting (Dohare et al., 2024). To analyze our method in this perspective, we visualize the cumulative success rates and changes in the L1 weight magnitude on the validation unseen split of REVERIE in Figure 5. Here, we observe that without any regularization or with a simple scaling method, the policy encounters the loss of plasticity and results in a gradual performance drop. This corresponds to the changes in weight magnitude, where the two variants exhibits the largest scale. Unlike these methods, SGR stands out with the lowest scale in weight magnitude, resulting in a stable increase throughout iterations. This is attributed to its counterfactual reasoning strategy that addresses the non-stationary nature of the binary learning environment.

Catastrophic Forgetting Analysis. Preserving the trained knowledge is as much important as acquiring new knowledge. Table 6 reports the results on the validation seen split,

Table 6. Experiments on Catastrophic Forgetting.

Methods	Val Unseen (✓) → Val Seen (✗)					
	TL ↓	OSR↑	SR↑	SPL↑	RGS↑	RGSPL↑
DUET (Chen et al., 2022c)	13.86	73.86	71.15	63.94	57.41	51.44
+ FSTTA (Gao et al., 2024a)	13.40	73.16	71.78	64.18	57.05	51.18
+ FEEDTTA w/o reg.	19.43	74.91	73.72	61.96	59.24	50.02
+ FEEDTTA w/ GD	22.01	72.24	71.05	57.19	57.20	46.72
+ FEEDTTA w/ GS	20.11	72.80	71.68	58.90	57.55	47.72
+ FEEDTTA w/ SGR	18.45	76.04	73.86	62.74	59.31	49.93

Table 7. Comparison of Feedback Strategies

Feedback Strategy	SR	SPL	RGSPL
Distance-based (Dense)	63.25	42.89	28.46
Goal-based (Sparse)	66.49	45.38	30.75

re-evaluated after the TTA on the validation unseen split to measure catastrophic forgetting. First, our FEEDTTA, without gradient regularization, enhances the OSR, SR and RGS metric after adaptation on the validation unseen dataset. This suggests that, beyond adaptation to a specific domain, the proposed feedback-based RL framework broadly enhances navigation success. We interpret the increase in TL as representing the minimal additional exploration required to achieve navigation success. Furthermore, while GD and GS exhibit catastrophic forgetting, the proposed SGR rather brings substantial improvements in the success rates, strengthening the policy's generalizability as well as adaptability on specific domain.

5.5. Comparison with Different Feedback Strategies

The rationale behind choosing a simple binary episodic feedback mechanism stems from the practical limitations of the online test-time navigation environment: (1) Human involvement should be minimal, as following every navigation steps to provide rewards is infeasible in real-world environment; and (2) Reward systems used in offline learning (e.g. step-wise distance-based rewards) are infeasible at test-time, as we assume no access to ground-truth goal position or pre-defined maps. We empirically evaluate the efficiency of the feedback system by comparing our method with the step-wise distance-based reward system used in HAMT, where the feedback is defined as the reduction in distance to the target at each step. Additionally, if the agents successfully arrives at the goal positions, 2 is given as a success signal and otherwise -2 as a penalty. As we observe from Table 7, our binary episodic feedback surpasses the distance-based dense reward system, even without access to ground-truth information. This clearly demonstrates that the proposed feedback mechanism appears to be simple, yet efficient and effective in improving navigation performance.

6. Conclusion

In this work, we introduce FEEDTTA, an effective TTA paradigm for online vision-language navigation that leverages feedback-based reinforcement learning. The proposed adaptation strategy utilizing binary episodic feedback enables agents to dynamically interact with their external environment by providing them with a notion of success and failure. Additionally, we develop a gradient regularization method, SGR, to robustly alleviate non-stationarity during adaptation. Through extensive experiments on challenging VLN benchmarks, our FEEDTTA showcases its superiority not only in traditional metrics but also in the proposed ASR metric, which evaluates the sample-wise transition of results before and after adaptation.

Limitation and Future Work. While LLMs show substantial potential as a replacement for human feedback, concerns about their reliability as oracles remain unresolved. The disparity between the accuracy of human judgment and LLM judgment must be minimized to ensure safer real-world applications. Since this work successfully incorporated the concept of feedback-based RL for the VLN task, we propose exploring more advanced applications of LLMs as feedback oracles as a promising direction for future research.

Impact Statement

Although our FEEDTTA leads significant performance improvements, it does not guarantee perfect prediction across the diverse environment. Therefore, significant attention must be paid for rigorous verification processes, prior to integrating it into the embodied AI system.

Acknowledgement

This work was supported by Samsung Advanced Institute of Technology (SAIT, 50%), Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (International Collaborative Research and Global Talent Development for the Development of Copyright Management and Protection Technologies for Generative AI, RS-2024-00345025, 13%; Research on neural watermark technology for copyright protection of generative AI 3D content, RS-2024-00348469, 25%), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2025-00521602, 10%), Institute of Information & communications Technology Planning & Evaluation (IITP) & ITRC(Information Technology Research Center) grant funded by the Korea government(MSIT) (No.RS-2019-II190079, Artificial Intelligence Graduate School Program(Korea University), 1%; IITP-2025-RS-2024-00436857, 1%), and Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea)&Gwangju Metropolitan City.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

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

Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., and Van Den Hengel, A. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3674–3683, 2018.

Black, K., Janner, M., Du, Y., Kostrikov, I., and Levine, S. Training diffusion models with reinforcement learning. *arXiv preprint arXiv:2305.13301*, 2023.

Chen, J., Gao, C., Meng, E., Zhang, Q., and Liu, S. Reinforced structured state-evolution for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15450–15459, 2022a.

Chen, P., Ji, D., Lin, K., Zeng, R., Li, T., Tan, M., and Gan, C. Weakly-supervised multi-granularity map learning for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 35:38149–38161, 2022b.

Chen, S., Guhur, P.-L., Schmid, C., and Laptev, I. History aware multimodal transformer for vision-and-language navigation. *Advances in neural information processing systems*, 34:5834–5847, 2021.

Chen, S., Guhur, P.-L., Tapaswi, M., Schmid, C., and Laptev, I. 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*, pp. 16537–16547, 2022c.

Dohare, S., Hernandez-Garcia, J. F., Lan, Q., Rahman, P., Mahmood, A. R., and Sutton, R. S. Loss of plasticity in deep continual learning. *Nature*, 632(8026):768–774, 2024.

Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.-P., Berg-Kirkpatrick, T., Saenko, K., Klein, D., and Darrell, T. Speaker-follower models for vision-and-language navigation. *Advances in neural information processing systems*, 31, 2018.

Gao, J., Yao, X., and Xu, C. Fast-slow test-time adaptation for online vision-and-language navigation. In *Forty-first International Conference on Machine Learning*, 2024a.

Gao, Z., Zhang, X.-Y., and Liu, C.-L. Unified entropy optimization for open-set test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23975–23984, 2024b.

Georgakis, G., Schmeckpeper, K., Wanchoo, K., Dan, S., Miltsakaki, E., Roth, D., and Daniilidis, K. Cross-modal map learning for vision and language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15460–15470, 2022.

Gu, J., Stefani, E., Wu, Q., Thomason, J., and Wang, X. E. Vision-and-language navigation: A survey of tasks, methods, and future directions. *arXiv preprint arXiv:2203.12667*, 2022.

Hao, W., Li, C., Li, X., Carin, L., and Gao, J. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13137–13146, 2020.

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

Irshad, M. Z., Mithun, N. C., Seymour, Z., Chiu, H.-P., Samarasekera, S., and Kumar, R. Semantically-aware spatio-temporal reasoning agent for vision-and-language navigation in continuous environments. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 4065–4071. IEEE, 2022.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Krantz, J., Wijmans, E., Majumdar, A., Batra, D., and Lee, S. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pp. 104–120. Springer, 2020.

Lee, H., Phatale, S., Mansoor, H., Lu, K. R., Mesnard, T., Ferret, J., Bishop, C., Hall, E., Carbune, V., and Rastogi, A. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023a.

Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., and Gu, S. S. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023b.

Li, X., Li, C., Xia, Q., Bisk, Y., Celikyilmaz, A., Gao, J., Smith, N., and Choi, Y. Robust navigation with language pretraining and stochastic sampling. *arXiv preprint arXiv:1909.02244*, 2019.

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

Liang, J., He, R., and Tan, T. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, pp. 1–34, 2024.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024a.

Liu, J., Xu, R., Yang, S., Zhang, R., Zhang, Q., Chen, Z., Guo, Y., and Zhang, S. Continual-mae: Adaptive distribution masked autoencoders for continual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28653–28663, 2024b.

Liu, J., Yang, S., Jia, P., Zhang, R., Lu, M., Guo, Y., Xue, W., and Zhang, S. Vida: Homeostatic visual domain adapter for continual test time adaptation. In *The Twelfth International Conference on Learning Representations*, 2024c.

Liu, R., Wang, W., and Yang, Y. Volumetric environment representation for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16317–16328, 2024d.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2025.

Long, Y., Li, X., Cai, W., and Dong, H. Discuss before moving: Visual language navigation via multi-expert discussions. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 17380–17387. IEEE, 2024.

Lu, Y., Zhang, H., Nie, P., Feng, W., Xu, W., Wang, X. E., and Wang, W. Y. Anticipating the unseen discrepancy for vision and language navigation. *arXiv preprint arXiv:2209.04725*, 2022.

Nado, Z., Padhy, S., Sculley, D., D'Amour, A., Lakshminarayanan, B., and Snoek, J. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020.

Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., and Tan, M. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pp. 16888–16905. PMLR, 2022.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Pashevich, A., Schmid, C., and Sun, C. Episodic transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15942–15952, 2021.

Peng, A., Netanyahu, A., Ho, M. K., Shu, T., Bobu, A., Shah, J., and Agrawal, P. Diagnosis, feedback, adaptation: A human-in-the-loop framework for test-time policy adaptation. In *International Conference on Machine Learning*, pp. 27630–27641. PMLR, 2023.

Pinto, A. S., Kolesnikov, A., Shi, Y., Beyer, L., and Zhai, X. Tuning computer vision models with task rewards. In *International Conference on Machine Learning*, pp. 33229–33239. PMLR, 2023.

Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W. Y., Shen, C., and Hengel, A. v. d. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9982–9991, 2020.

Qiao, Y., Qi, Y., Hong, Y., Yu, Z., Wang, P., and Wu, Q. 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*, pp. 15418–15427, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization:

Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., and Bethge, M. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*, 33: 11539–11551, 2020.

Tan, H., Yu, L., and Bansal, M. Learning to navigate unseen environments: Back translation with environmental dropout. *arXiv preprint arXiv:1904.04195*, 2019.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Tseng, H.-Y., Chen, Y.-W., Tsai, Y.-H., Liu, S., Lin, Y.-Y., and Yang, M.-H. Regularizing meta-learning via gradient dropout. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Wang, D., Shelhamer, E., Liu, S., Olshausen, B., and Darrell, T. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020a.

Wang, H., Wu, Q., and Shen, C. Soft expert reward learning for vision-and-language navigation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pp. 126–141. Springer, 2020b.

Wang, Q., Fink, O., Van Gool, L., and Dai, D. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022.

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

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

Wu, W., Chang, T., Li, X., Yin, Q., and Hu, Y. Vision-language navigation: a survey and taxonomy. *Neural Computing and Applications*, 36(7):3291–3316, 2024.

You, F., Li, J., and Zhao, Z. Test-time batch statistics calibration for covariate shift. *arXiv preprint arXiv:2110.04065*, 2021.

Yu, B., Kasaei, H., and Cao, M. Co-navgpt: Multi-robot cooperative visual semantic navigation using large language models. *arXiv preprint arXiv:2310.07937*, 2023.

Zhang, M., Levine, S., and Finn, C. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35:38629–38642, 2022.

Zhao, B., Chen, C., and Xia, S.-T. Delta: degradation-free fully test-time adaptation. *arXiv preprint arXiv:2301.13018*, 2023a.

Zhao, S., Wang, X., Zhu, L., and Yang, Y. Test-time adaptation with clip reward for zero-shot generalization in vision-language models. In *The Twelfth International Conference on Learning Representations*, 2023b.

Zheng, D., Huang, S., Zhao, L., Zhong, Y., and Wang, L. Towards learning a generalist model for embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13624–13634, 2024.

Zhou, G., Hong, Y., and Wu, Q. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7641–7649, 2024.

Zhou, G., Hong, Y., Wang, Z., Wang, X. E., and Wu, Q. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In *European Conference on Computer Vision*, pp. 260–278. Springer, 2025.

A. Details of Large Language Model Oracles

In Figure 6, we provide the prompts we utilized for the experiments in a dialogue format. As mentioned in Sec 5.3, the process leverages a two-step architecture, where the LLM first identifies the target goal from the given instruction, and then determines the navigation success/failure based on the alignment of the goal with the panoramic image from the last navigation step. The system offers an effective solution in scenarios where human feedback is unavailable.

Example dialogue for the REVERIE dataset

```
You are an indoor navigation robot executing the following instruction.
Instruction: {instruction_txt}
Task: Considering this instruction, what is the place or object you should ultimately look for?
Answer in simple word or phrase. Do not include any additional text in your response.
```

Answer : {response_goal}

```
Your Task:
- Analyze the given panoramic view image together to identify if the image generally describe a
scenario where the {response_goal} is visible.
- Conclude whether the navigation is a success or failure based on a broad alignment between the
instruction and image.

Evaluation Criteria:
1. Yes: If the {response_goal} can be reasonably found with the given image.
2. No: If the {response_goal} is inconsistent with the given image, with no reasonable alignment
or evidence.

Important Notes:
1. Use reasonable inference to find the {response_goal} in the given image. If the image broadly
describe the target object(e.g., "armchair" can be replaced with "chair"), treat it as a success.
2. Partial Alignment: If the majority of the details in the given image align with the instruction,
consider it a "Yes".
Exact positioning or minor missing details should not override a clear overall match.

Input:
- natural language instruction: {instruction_txt}
- Panoramic Image:
```

```
Output:
- Your response should only be "Yes" or "No".
```

Answer : “Yes” or “No”

Figure 6. **Overall pipeline of LLMs as an oracle.**

B. Details of Stochastic Gradient Reversion

B.1. Derivation of the Scaling Factor in Eq.5

We provide a mathematical derivation of how the scaling factor of $\frac{1}{\alpha p+(1-p)}$ in Eq. 5 can ensure consistency in expectation.

Step 1. Expectation of the Modified Gradient: The modified gradient can be written as:

$$g'_{\theta_m} = g_{\theta_m} \cdot (\alpha \cdot b_m + (1 - b_m)).$$

Taking the expectation over b_m, where $\mathbb{E}[b_m] = p$, we have:

$$\mathbb{E}[g'_{\theta_m}] = g_{\theta_m} \cdot \mathbb{E}[\alpha \cdot b_m + (1 - b_m)].$$

Substituting $\mathbb{E}[b_m] = p$, the expectation becomes:

$$\mathbb{E}[g'_{\theta_m}] = g_{\theta_m} \cdot (\alpha p + (1 - p)).$$

Step 2. Scaling the Modified Gradient: To ensure consistency in expectation, we scale g'_{θ_m} by $\frac{1}{\alpha p + (1-p)}$. The scaled gradient is:

$$\tilde{g}'_{\theta_m} = \frac{g'_{\theta_m}}{\alpha p + (1 - p)}.$$

Step 3. Expectation of the Scaled Gradient: Taking the expectation of $\tilde{g}'_{\theta_m}$, we get:

$$\mathbb{E}[\tilde{g}'_{\theta_m}] = \mathbb{E}\left[\frac{g'_{\theta_m}}{\alpha p + (1 - p)}\right] = \frac{\mathbb{E}[g'_{\theta_m}]}{\alpha p + (1 - p)}.$$

From Step 1, $\mathbb{E}[g'_{\theta_m}] = g_{\theta_m} \cdot (\alpha p + (1 - p))$. Substituting this gives us:

$$\mathbb{E}[\tilde{g}'_{\theta_m}] = \frac{g_{\theta_m} \cdot (\alpha p + (1 - p))}{\alpha p + (1 - p)} = g_{\theta_m}.$$

B.2. Reversion Rate p and Reversion Magnitude α

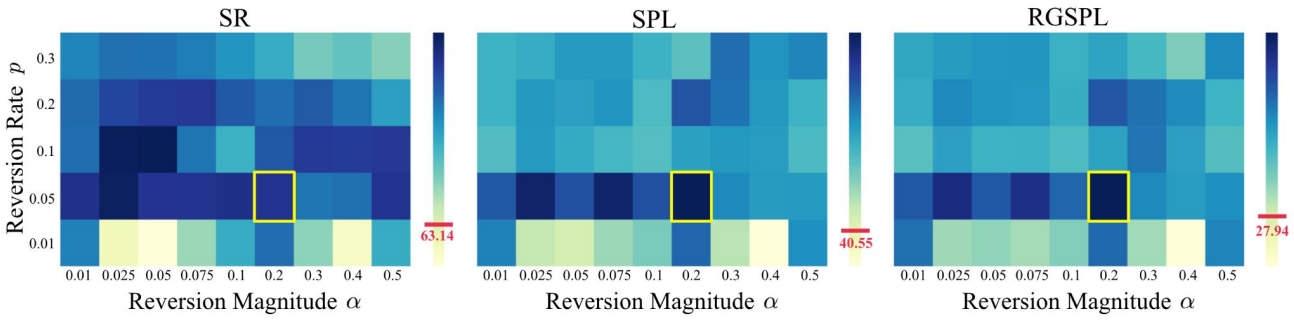

Figure 7. **Hyperparameter Analysis of SGR.** We illustrate the performance variations w.r.t. the reversion rate p and the reversion magnitude α. The red markers in the color bar indicate the performances of DUET w/ FEEDTTA before applying the SGR regularization and the yellow-boxed cells are the reported combination throughout the manuscript.

In this section, we analyze the performance variations for the combinations of the reversion rate p and the reversion magnitude α in SGR. Figure 7 illustrates the results of FEEDTTA on the validation unseen split of REVERIE, measured on three metrics SR, SPL, and RGSPL. DUET (Chen et al., 2022c) is utilized as the target policy. Here, we observe three insights: (1) SGR generally brings improvements in the performance, demonstrating its robustness on various configurations of p and α. (2) A reversion rate of $p = 0.05$ generally yields decent navigation performances. (3) Regularizing an excessively small number of parameters with $p = 0.01$ has minimal effect on performance. (4) Reversing gradients with magnitudes exceeding 0.3 results in lower SPL and RGSPL, suggesting increased exploration during navigation.

C. Details of Evaluation Metrics

Below, we provide the details of the evaluation metrics that are used throughout our experiments.

- **Trajectory Length (TL)** measures the average distance the agent traveled from the starting point to the endpoint in metric units. Lower value typically indicates efficient navigation.

Table 8. General TTA sequence ordering.

Method	SR	SPL	RGSPL
DUET	46.98	33.73	23.03
w/ FEEDTTA	65.33 (±1.10)	42.63 (±1.98)	28.71 (±1.45)

Table 9. Continual TTA sequence ordering.

Method	SR	SPL	RGSPL
DUET	46.98	33.73	23.03
w/ FEEDTTA	54.81 (±1.89)	36.70 (±0.87)	23.74 (±0.47)

Table 10. Per-Scene TTA sequence ordering

Scene ID	1	2	3	4	5	6	7	8	9	10
SR	48.63 / 60.78 (±1.65)	72.22 / 77.16 (±0.87)	32.65 / 37.51 (±0.96)	46.85 / 51.75 (±1.00)	43.34 / 51.09 (±3.25)	30.30 / 39.46 (±4.49)	44.84 / 56.76 (±5.92)	50.60 / 64.57 (±7.48)	45.89 / 71.54 (±2.73)	55.67 / 69.47 (±1.53)
SPL	30.74 / 40.80 (±0.96)	56.61 / 61.51 (±1.73)	19.88 / 22.79 (±0.49)	37.27 / 38.98 (±1.30)	25.55 / 29.31 (±3.14)	19.81 / 24.73 (±3.57)	35.21 / 30.53 (±3.12)	34.36 / 38.59 (±5.51)	29.54 / 50.29 (±6.74)	44.91 / 56.42 (±2.87)

- **Navigation Error (NE)** measures the average distance between the ground truth endpoint and the predicted endpoint in metric units. Lower value indicates that the agent closely followed the given instruction.
- **Success Rate (SR)** calculates the fraction of successful navigation over total navigation attempts, where NE < 3 is considered as a success.
- **Oracle Success Rate (OSR)** calculates the fraction of successful navigation over total navigation attempts, where it is considered as a a success if one of the navigation points in the trajectory contains a ground truth endpoint.
- **Success penalized by Path Length (SPL)** evaluates the weighed trajectory efficiency for navigation success, where a score closer to SR indicates that the trajectory closely followed the shortest path. The equation is formulated as $\text{SPL} = \frac{1}{N}\sum_{n=1}^{N} S_n \frac{\text{TL}_n}{\max(\text{SP}_n, \text{TL}_n)}$, where S is the binary indicator for success and SP denotes the shortest path.
- **Remote Grounding Success (RGS)** measures the portion of navigation attempts that successfully grounded the target object required from the instruction, determined by a bounding box prediction with IoU (intersection over union) ≥ 0.5 compared to the ground truth.
- **Remote Grounding Success penalized by Path Length (RGSPL)** calculates the weighed trajectory efficiency for remote grounding success, similar to the SPL metric.

D. Effects of Different Sequence Ordering

Online learning is inherently sequence-dependent. However, we show that the benefits of FEEDTTA is invariant to sequence ordering through the following experiments with three different configurations. We use the 'validation unseen' split of the REVERIE dataset and compare with the DUET policy. For all configurations, the reported numbers of FEEDTTA are the average of the results from 3 different seeds, with standard deviation reported in brackets.

- **General TTA:** In this configuration, all episodes are randomly ordered regardless of scene IDs, which corresponds to the experimental setting reported in Table 1 of our paper. The results are reported in Table 8.
- **Continual TTA:** For this configuration, we fix the episode orders for each scene ID, and set the adaptation sequence based on mixed scene ID orders, evaluating continual adaptation performances across different scenes. The results are reported in Table 9.
- **Per-Scene TTA:** Here, we analyze the effects of random episode orders for each scene ID. Note that in this setting, the adaptation is performed per-scene, and not throughout the entire validation set. The results are reported in Table 10 in the form of (DUET / +FeedTTA).

These experiments confirm that sequence ordering does influence navigation outcomes; however, the benefits of FEEDTTA remain consistent, as evidenced by superior performances with low variations across different seeds.

E. Trajectory Visualization

In this section, we analyze the trajectories of DUET before and after applying our FEEDTTA with visual illustrations. For this study, we select two episodes that initially fail under the base DUET policy but achieve success following a one-step parameter update with our FEEDTTA. The visualized results in Figure 8 provides the following insights.

(1) The base DUET policy navigates to a destination that closely resembles the instructed location but often fails due to missing critical details. For instance, in the first sample shown on the left of Figure 8, the agent following the DUET policy navigates to the bathroom, as specified by the instruction. However, the predicted destination does not fully align with the specific details provided in the instruction. This can be also observed in the second example, where the agent stops near the kitchen near the family room, but fails to reach the precise location described in the instruction.

(2) Upon receiving negative feedback via FEEDTTA after a navigation failure, the agent adjusts its trajectory at a certain point that we refer to as an *uncertainty zone*. Uncertainty zone is a visual state with multiple feasible navigable locations, resulting in high uncertainty (*i.e.*, entropy). In both examples, the measured entropy at the uncertainty zone is the highest observed along the DUET trajectory. The updated policy from FEEDTTA chooses an alternative path within the uncertainty zone, enabling exploration of new possibilities.

(3) Only with a single-step parameter update with feedback, the agent successfully arrives at the desired destination. Multiple iterations from the online data stream further enhance adaptability, and generalizability by learning the notion of success and failure, which is demonstrated throughout our experiments.

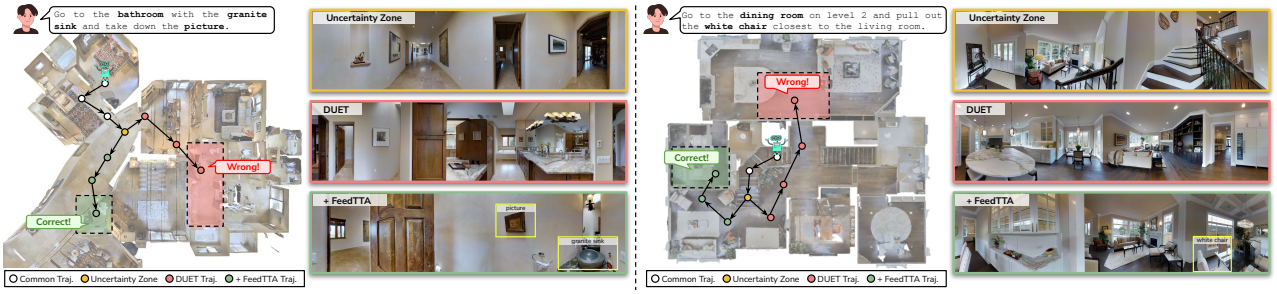

Figure 8. **Visual Analysis of FEEDTTA.** We illustrate two examples of episodes that initially fail under the base policy but achieve success after applying FEEDTTA. We provide a top-down view of the trajectories, along with a panoramic image of the uncertainty zone and the endpoints reached by DUET and FEEDTTA.

F. FAQ & Discussions

In this section, we share some of the notable questions and discussions that emerged throughout the research process.

Question 1: How is learning from online feedback different from learning from ground truths?
Answer: Online feedback and ground truth are inherently different in the VLN task. First, ground truth consists of offline collected state-action pairs for each step, whereas online feedback can be any scalar value based on the oracle's objective. Accordingly, ground truths directly force the policy to learn optimal action, which guarantees performance when strictly followed. However, feedback provides indirect guidance by encouraging actions that maximize the reward, which may vary depending on the oracle's preferences. Considering these aspects, learning from ground truths is infeasible and impractical in the online TTA setup, whereas feedback-based learning offers a more adaptable framework by enabling policies to iteratively improve through interaction with the environment and alignment with the oracle's objectives.

Question 2: Why is the performance improvement smaller in the R2R datasets compared to that of REVERIE's?
Answer: We assume this is due to differences in the instructions between the two datasets and their alignment with the binary episodic feedback mechanism of FEEDTTA. While FEEDTTA demonstrates its superior test-time adaptability in both datasets, the R2R-trained policy and the REVERIE-trained policy inherently require different guidance for performance improvements. Given that the former relies on dense, step-wise guidance during training, the binary episodic feedback provided by FEEDTTA might be relatively sparse to drive significant performance enhancements. However, the latter, which is trained to operate under less structured and more abstract instructions, is better suited to benefit from the sparse binary episodic feedback of FEEDTTA, allowing it to adapt more effectively during test-time.

Question 3: Can FEEDTTA adapt to novel navigation tasks at test time?
Answer: Continuing from the previous analysis, we empirically address this question by conducting experiments on the R2R dataset (*i.e.*, step-by-step instruction following task) with a policy trained on the REVERIE dataset (*i.e.*, goal-oriented task), and similarly, on the REVERIE dataset using a policy trained on R2R. Evaluations are carried out on the validation unseen split of both dataset. Table 11 shows that while the absence of fine-grained trajectory details in REVERIE instructions leads

to longer TL for R2R-trained policies to identify the goal point, both training brings improvement in the navigation success of each task. This suggests that FEEDTTA effectively leverages the shared knowledge underlying both tasks, which involves reaching a target point based on given instructions. However, the results from the RGSPL metric suggest that adapting and improving on an untrained task in real-time using a single-stream online learning approach is challenging.

Table 11. Experimental results on the cross-task TTA.

REVERIE→R2R	NE ↓	SR	SPL	**R2R→REVERIE**	TL ↓	SR	RGSPL
DUET (Chen et al., 2022c)	9.78	17.33	4.82	DUET	14.67	24.91	3.47
w/ FEEDTTA	8.94	18.35	6.51	w/ FEEDTTA	22.50	28.86	3.49

Question 4: Can LLM-generated counterfactual evaluations replace SGR?
Answer: The counterfactual reasoning of SGR is a regularization technique applied on a limited number of parameters, which means that the large portion of parameters should be updated based on proper feedback for intended functionality. Furthermore, while LLMs can indeed reason counterfactual scenarios, their reliability on predicting navigation outcomes itself still remains as a challenge, making them unsuitable as a direct replacement for SGR.

Question 5: Can FeedTTA be applied to Visual Navigation tasks?
Answer: Yes, FEEDTTA can be applied to Visual Navigation (VN) tasks even in the absence of complex language instructions, as it only requires determining success or failure within the navigation system. To identify the dominant modality influencing navigation outcomes, we analyze navigation consistency for each trajectory in the REVERIE dataset, where each trajectory is paired with multiple language instructions. Specifically, we compute the average success rate across different instructions for each trajectory. We then identify trajectories with consistent outcomes—defined as those with a high (> 0.8) or low (< 0.2) average success rate—and calculate their proportion within the validation set. Our experiment yields a ratio of 0.72, suggesting that visual observations are a key factor not only in VN tasks but also in VLN, where they play a more decisive role compared to language variations.

Question 6: Does increased trajectory length in some episodes represent beneficial exploration?
Answer: We justify that the increased trajectory length (TL) indicates beneficial exploration by empirically testing the hypothesis: *"The overall increase in TL primarily results from episodes that would have failed in the original navigation but succeeded after applying FeedTTA"*. In Table 12, we compare the increase in TLs for the successful navigation episodes after adaptation, categorized based on the pre-tested results before applying FEEDTTA. For this experiment, we use the 'validation unseen' split of the REVERIE dataset with DUET as the base policy. Here, we discover that the average TL increase is significantly larger for fail-to-success cases than for success-to-success cases. This clearly demonstrates the important role of FEEDTTA in overcoming failure cases through extended exploration in unseen navigation environment.

Table 12. Trajectory length and exploration

	Success→Success	**Fail→Success**
Increased TL	3.54 ($\pm$1.45)	10.65 ($\pm$3.75)

Question 6: What new directions for research does FEEDTTA suggest?
Answer: FEEDTTA serves as an exemplar of incorporating the recent advancements of the feedback-based RL into the robotic navigation task. Considering the contributions and limitations of FEEDTTA, we suggest the following topics as prospective future research directions:

- **Advanced application of LLMs as navigation oracles.** As highlighted in Sec.5.3 and Sec.6, enhancing the accuracy of navigation outcome predictions is essential to ensure safer deployment of LLMs as navigation oracles. One approach is to develop a more advanced LLM architecture and prompting system capable of capturing the complex reasoning underlying its predictions. Another way is to incorporate visual foundation models (Radford et al., 2021; Kirillov et al., 2023; Liu et al., 2025) to provide LLMs with more spatial contexts. Both approaches will enhance the reliability of LLMs as oracles, benefiting not only the TTA of VLN but also zero-shot VLNs (Zhou et al., 2024; Long et al., 2024).
- **Test-time Adaptation on untrained navigation tasks.** In real-world scenarios, the given instructions and tasks may differ from the trained navigation tasks, leading to the outcomes shown in Table 11. Therefore, it is crucial to develop a generalized TTA for diverse navigation tasks to ensure the versatility of embodied agents in real-world applications.
- **Feedback-based RL for offline VLN training.** RL is accommodated in several previous literature (Chen et al., 2021; 2022a) with heuristic reward shaping. However, with the success of feedback-based RL in online navigation with

FEEDTTA, training the notion of success and failure through binary episodic feedback can alleviate the burden of manual reward shaping. We speculate that combining policies trained with offline feedback-based RL and online TTA using feedback-based RL techniques, such as FEEDTTA, could yield substantial synergy.