DialNav: Multi-turn Dialog Navigation with a Remote Guide

Anonymous ICCV submission

Paper ID

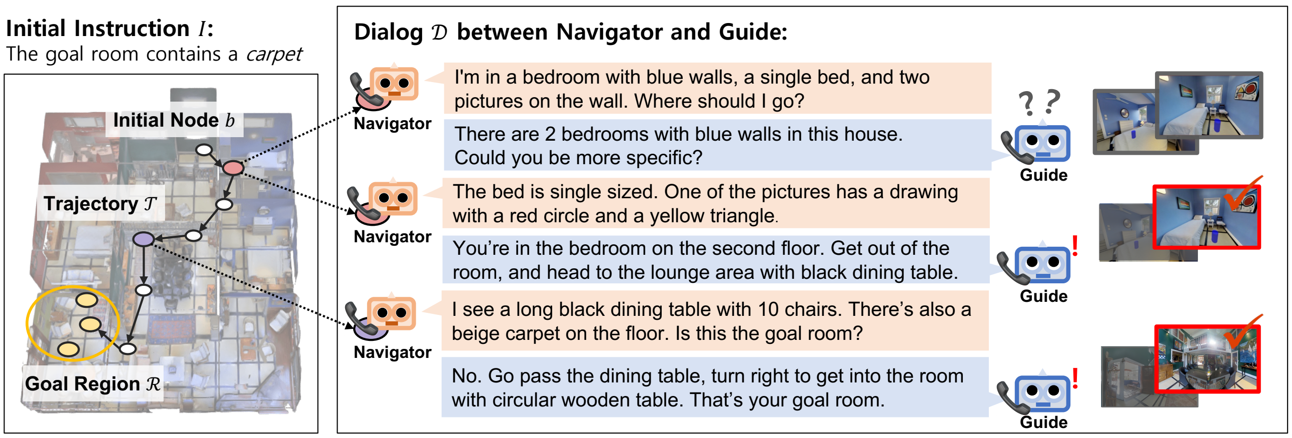

Figure 1. **Overview of the proposed DialNav task.** Navigator is tasked with reaching the goal region $\mathcal{R}$ (yellow circle in the map) from the initial node b based on an ambiguous initial instruction I, which provides only a hint about $\mathcal{R}$ (*e.g.*, '*The goal room contains a carpet*'). During navigation, Navigator can ask questions to obtain additional guidance (orange text boxes). Guide has knowledge of the environment but lacks information of Navigator's location. For successful navigation, Guide must estimate Navigator's location through dialog $\mathcal{D}$ and provide directions from the estimated position to the goal region (blue text boxes). Note that each QA pair in $\mathcal{D}$ is mapped to a node (red and purple nodes mapped to dialog turns by dotted arrows) in the navigation trajectory $\mathcal{T}$.

Abstract

We introduce DialNav, a novel collaborative embodied dialog task, where a navigation agent (Navigator) and a remote guide (Guide) engage in multi-turn dialog to reach a goal location. Unlike prior work, DialNav aims for holistic evaluation and requires the Guide to infer the Navigator's location, making communication essential for task success. To support this task, we collect and release Remote Assistance in Navigation (RAIN) dataset, human-human dialog paired with navigation trajectories in photorealistic environments. We design a comprehensive benchmark to evaluate both navigation and dialog, and conduct extensive experiments analyzing the impact of different Navigator and Guide models. We highlight key challenges and publicly release the dataset, code, and evaluation framework to foster future research in embodied dialog. Our code and dataset are available at: https://happilee12.github.io/DialNav.

1. Introduction

An embodied AI agent with a physical body perceives and interacts with the world, responding to human instructions or environmental stimuli. Since an embodied agent physically engages with its surroundings, misinterpreting human commands or hallucinating unintended actions can cause inconvenience or even physical harm. To mitigate these risks, the agent must seek clarification when faced with ambiguous tasks. A natural approach is dialog-based interaction, allowing the agent to refine its understanding before acting. This enhances both safety and effectiveness in task execution.

Despite the importance of communicative ability in embodied AI, its inherent challenges hinders progress. First, collecting dialog-based data is costly, as it requires two individuals to engage in real-time interaction while performing the task. Secondly, a dialog-based task necessitates not only equipping the agent with the ability to ask questions but also constructing a counterpart model capable of providing responses. Additionally, a framework that enables task execution through question-and-answer interactions is required.

Lastly, the dynamic and interdependent nature of the task makes it difficult to evaluate performance. Due to these challenges, prior research has predominantly focused on task execution rather than leveraging dialog for task completion.

Some studies have already acknowledged the importance of dialog in embodied AI and has explored tasks involving dynamic additional instructions from a guide [13, 23–25, 31]. However, these studies often assume that the guide has full knowledge of current situation—an unrealistic assumption in practical settings. This omniscient guide model reduces the incentive for the guide to carefully consider the performer's questions and, in turn, discourages the performer from formulating high-quality inquiries. Even a vague request like "help" can elicit a perfect response from an all-knowing guide, making the dialog less meaningful.

To address this limitation, we propose DialNav. DialNav is a cooperative navigation task between Navigator and Remote Guide, where Guide is not aware of Navigator's location but only familiar with the environment. This setting closely mirrors real-world scenarios, such as when a lost person calls a friend for directions. Here, Navigator must formulate high-quality, detailed questions, as the effectiveness of Guide's guidance depends on the clarity of the queries. Likewise, Guide is incentivized to pay close attention to Navigator's questions, as they lack direct awareness of Navigator's location. Some prior works [2, 9] have explored non-omniscient guides, but under restricted conditions—[2] evaluates agents using static dialog histories in a physical maze, while [9] limits the guide's view to a 2D grid map, reducing spatial fidelity. In contrast, DialNav supports long-horizon dialog and navigation in realistic indoor environments, enabling richer and more purposeful communication.

To support this task, we introduce the Remote Assistance in Navigation dataset (RAIN), a visual navigation dataset with multi-turn human-human dialog. To this end, we developed a data collection tool, specifically designed to capture complex interactions in DialNav. As RAIN dataset can be used to train and evaluate the core capabilities of Navigator and Guide under standardized conditions, we conduct experiments with Navigator and Guide models across various setups and module configurations, providing valuable insights into this new task. We will publicly release the complete codebase, data collection tool, dataset, and training and evaluation frameworks, ensuring reproducibility and providing foundational resources for future research. Our main contributions are summarized as follows:

- We introduce a cooperative embodied dialog navigation task involving Navigator and remote Guide.
- We collect and release RAIN, which captures human-human interactions for DialNav.
- Navigator and Guide are trained and tested under various setups within a holistic dialog-and-navigation pipeline.
- We discuss key challenges and their implications for fostering future research.

2. Related Work

Embodied Dialog Recent advancements in embodied dialog have led to the development of new benchmarks and datasets across question answering [7, 21], navigation [10, 31], and manipulation [13, 22, 25, 30]. Although these benchmarks incorporate dialog, they either assume an omniscient guide [13, 22, 23, 25, 28] or restrict evaluation to task execution only [2, 15, 31], thereby underestimating the role of dialog (*e.g.*, asking for help without a question [23]). Our work models two dialog-enabled agents, Navigator and remote Guide, and emphasizes the importance of dialog generation and understanding within the full communication-action loop. The most closely related work is Talk the Walk [9] but it operates in a highly constrained setting with limited agent actions and simplified 2D grid map for the guide. In contrast, our task supports long-horizon, multi-turn dialog in a realistic navigation environments.

Vision and Language Navigation Vision-and-Language Navigation (VLN) is a prominent multimodal embodied task where an agent navigates a visual environment based on language instructions. Anderson et al. [1] pioneered the field by introducing the first photo-realistic VLN task, dataset, and a simulator. Since then, a wide range of VLN benchmarks have emerged, but they often face limitations, either providing overly detailed instructions [17, 18] unnatural for human interaction or offering ambiguous guidance [27, 40] insufficient for successful navigation. To overcome these limitations, this study explores a realistic dialog-based navigation task.

Subtasks for DialNav DialNav targets the holistic evaluation of the communication-and-navigation process, integrating multiple subtasks that have previously been addressed only in isolation. Instruction generation [4, 11, 12, 33, 35, 36, 38, 41], explored primarily for data augmentation in VLN, is closely related to the question and answer generation components of DialNav. Navigator localization is a distinguishing aspect of DialNav compared to previous VDN tasks. Unlike SLAM [8] focusing on self-localization, this task identifies the agent location based on natural language description. The Localization from Embodied Dialog (LED) task was introduced in [15], followed by a few subsequent studies [14, 19, 26, 34, 39]. While navigator localization in DialNav shares similarities with LED, it presents additional challenges due to its dynamic navigation context and long-form dialog. Deciding when to ask for help has been explored in prior work [13, 29, 42], often by introducing an auxiliary model or leveraging action space entropy. We adopt these models for the corresponding modules in DialNav.

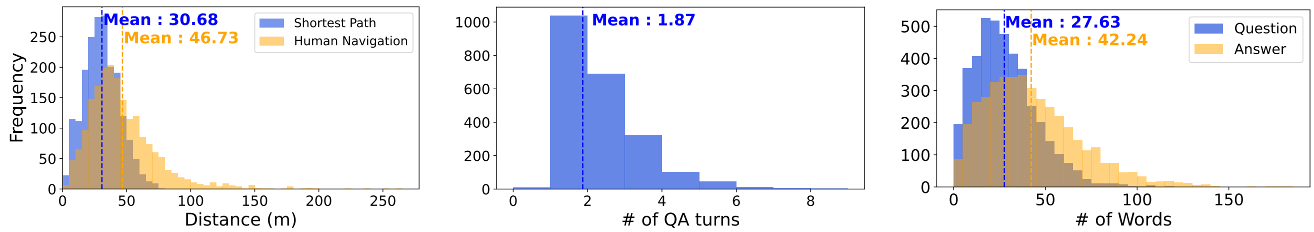

Figure 2. **Data distributions in RAIN.** Distributions of trajectory lengths for shortest paths and human navigation (left), QA pairs per episode (center), and word counts per question and answer (right).

Tasks	Subtasks					Guide
	N	*W*	*Q*	*L*	*A*	
VLN [1]	✓					✗
Just Ask [6]	✓	✓				Omniscient
HANNA [23]	✓	✓				Omniscient
VDN [31]	✓	✓	✓		✓	Omniscient
DialNav(Ours)	✓	✓	✓	✓	✓	Remote

Table 1. **Comparing vision-and-language navigation tasks in photo-realistic environments.** The 'Subtasks' column indicates tasks involved in DialNav: *N*: navigation, *W*: deciding whether to ask, *Q*: question generation, *L*: localizating navigator, *A*: answer generation. Navigation, deciding whether to ask, and question generation are subtasks for Navigator, while localizing navigator and answer generation are subtasks for Guide.

3. Task and Dataset

3.1. The DialNav Task

DialNav is a novel navigation task that targets the holistic evaluation of the dialog and execution loop, where an agent (*Navigator*) navigates to a specified goal region with assistance from a remote guide (*Guide*). Unlike prior studies [13, 23–25, 28, 29], DialNav incorporates a remote guide and holistic evaluation, encompassing all key subtasks for dialog-enabled navigation.

In DialNav, Navigator starts with an ambiguous instruction and gathers additional information through natural language dialog with Guide. Navigator must decide when to seek assistance and generate relevant questions. Guide, upon receiving a query, infers Navigator's location before providing a response, incentivizing Navigator to ask informative questions. Dialog is initiated by Navigator and follows an alternating turn-taking format, with no limit on QA turns. Tab. 1 compares DialNav to existing VLN tasks in photo-realistic environments.

3.2. Dataset Collection

To address DialNav, we collected RAIN, a dataset consisting of episodes in which two human annotators interact to complete a navigation task. This section outlines the dataset curation process and presents a statistical analysis of its key attributes.

Navigation Simulator Our study is based on the Matterport3D simulator [1], which utilizes photo-realistic reconstructions of real-world houses. The simulator represents each reconstructed house as a graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ and $\mathcal{E}$ denote a set of navigable nodes representing locations within the house and a set of edges connecting nodes, respectively. An agent can traverse over the environment graph G moving between nodes $v \in \mathcal{V}$ connected by edges $e \in \mathcal{E}$. The agent is given a panoramic view at each node. We use 83 house scans partitioned into 61 for train and val seen, 11 for val unseen and 18 for test respectively following [1]. The distinction between val seen and val unseen is based on whether the house is included in the training split.

Episodes RAIN contains a total of 2,231 DialNav episodes. Each episode $E = (G, b, \mathcal{R}, I, \mathcal{T}, \mathcal{D})$ is a 6-tuple composed of an environment graph G, Navigator's initial node $b \in \mathcal{V}$, a set of nodes $\mathcal{R} \subset \mathcal{V}$ in the goal region, an initial instruction I, Navigator's navigation trajectory $\mathcal{T}$ from b to one of goal nodes $r \in \mathcal{R}$, and the dialog $\mathcal{D}$ between Navigator and Guide. Note that the first four elements $(G, b, \mathcal{R}, I)$ define a navigation task and a goal region (*e.g. living room* or *bathroom*) may contain multiple nodes ($|\mathcal{R}| \geq 1$). Our dataset includes 1,401 navigation tasks from CVDN [31] and 838 additional tasks generated following the same scheme. The trajectory $\mathcal{T}$ and dialog $\mathcal{D}$ are collected from human annotators. A trajectory $\mathcal{T} = (v_1, v_2, ..., v_N)$ is a sequence of nodes $v_i \in \mathcal{V}$ with $v_1 = b$ and $v_N \in \mathcal{R}$. A dialog $\mathcal{D} = (d_1, d_2, ..., d_M)$ where $d_i = (q_i, a_i, u_i)$ is a triplet of a question q_i posed by Navigator, its corresponding answer a_i from Guide and the node u_i where this question-answer exchange occurs. The node u_i represents the last visited node in the trajectory at the corresponding dialog turn.

Data Collection Tool We developed a data collection tool to simulate DialNav. It provides Navigator with a navigation interface and a real-time chat interface for communication with Guide. Meanwhile, unlike previous studies that provide the guide with next optimal steps from Navigator's exact location, our tool requires Guide to infer Navigator's position using a specialized interface that displays the entire house layout and room list. Once Guide selects a node, the tool provides the shortest path trajectory to the goal from that node, assisting in answer generation. For each episode, two annotators are randomly assigned as Navigator and Guide, and given a task $(G, b, \mathcal{R}, I)$. More details about our tool can be found in Sec. H in Supp. Mat. We will release the

Dialog characteristics	Init.	Subs.	Example
Nav describes scene	0.97	0.74	*Nav*: **I'm the living room which is connected to the kitchen area. There's two beige arm chairs and one beige soft with beige and blue colored cushions on it. ...**
Gui describes path	0.86	0.75	*Gui*: **go upstairs; turn right once you're upstairs and go straight along the hallway. To your left, there should be an open door with a circular table on top of a large carpet.** ...
Gui requests clarification	0.15	0.03	*Gui*: By any chance, **do you see a swimming pool?**
Nav requests clarification	-	0.13	*Nav*: sharp left turn, **you mean go through between the stove and wooden big table ?**
Confirms goal	-	0.14	*Nav*: ... there is a statue of woman in front of the long wooden table. **Am I at the goal room?** *Gui*: **Yes correct.** good work
Need dialog history	-	0.46	*Nav*: yea now i'm standing in front of **those two statues**

Table 2. **Dialog characteristics in RAIN.** We manually analyzed 100 randomly sampled episodes from RAIN and present various dialog characteristics in RAIN, along with an example. The Init. and Subs. columns indicate the frequencies of these characteristics in the first and subsequent dialog turns, respectively.

Q: Think I'm at the room what you've noticed me. On my left side, there is a table made of transparent glass and on my right side, there is a Queen side bed. Also with the two doors opened to the balcony.

A: That's right. Go out the left door to the balcony, you'll see two black stools. That balcony is your target room.

(a) RAIN

Q: Should I exit this room?

A: Yes, go out blue door.

(b) CVDN

Figure 3. **Example QA pairs from RAIN and CVDN.** RAIN questions include detailed environmental descriptions to help the remote guide infer Navigator's location. In contrast, CVDN questions simply request the next action, as detailed descriptions are unnecessary for an omniscient guide. Additional examples from RAIN are provided in Supp. Mat.

code for the collection tool to support future research.

Data Collection Process To ensure smooth data collection, annotators first watched a tutorial video and completed two practice episodes beforehand. All annotators provided informed consent for data usage in research. To ensure high-quality dataset, human annotators evaluated each other on a 5-point scale after completing each episode. The average scores are 4.48 and 4.28 for Navigator and Guide respectively. We imposed a time limit of 22 minutes for each episode, ensuring that only efficient plays were included. One episode took 8 minutes in average. The final dataset comprises 2,231 episodes, distributed across training (1,559), validation seen (111), validation unseen (276), and test (285) sets. The total cost for data collection was approximately 7,500 USD. Refer to Sec. I in Supp. Mat. for further details.

3.3. Statistics

Trajectory Fig. 2(left) compares trajectory length distributions between shortest paths and human navigation in RAIN. Shortest paths range from 2.87m to 76.55m (avg. 30.68m with 17.39 nodes), while human navigators travel 3.02m to 262.64m (avg. 46.73m with 25.97 nodes), making human trajectories 1.62 times longer on average. The distribution is long-tailed, with over 80% deviating less than twice the shortest path length. The largest deviation is 33.5 times longer than the shortest path. Large deviations in general are attributed to Navigator's excessive exploration, especially when the ground-truth trajectory is short.

Dialog Fig. 2(center) presents the distribution of QA turns in RAIN. Each episode contains an average of 1.87 QA pairs, with over 92% concluding within three QA pairs. Ten episodes include no dialog, while the longest interaction consists of eight QA pairs. As shown in Fig. 2(right), questions and answers in RAIN average 27.63 and 42.24 words, respectively. Since Guide has extensive knowledge of the environment and each QA turn requires localizing Navigator, which is a costly process, Guide tends to provide comprehensive responses when possible, leading to fewer QA turns but longer answers.

RAIN Dialog Features Tab. 2 highlights key dialog features in RAIN. DialNav demands detailed dialog and continuous verification to ensure Navigator and Guide stay aligned. Navigator provides rich descriptions of its surroundings when asking questions (Row 1), while Guide sometimes seeks clarification for unclear queries (Row 3), emphasizing the importance of precise questions in DialNav. This contrasts with CVDN [31], where Guide gives directions based solely on perfect knowledge of Navigator's location (Fig. 3). Additionally, due to the ambiguous initial instruction in DialNav, Navigator and Guide engage in extra dialog upon reaching the goal for confirmation (Row 5).

4. Navigator and Guide Models

The collaborative nature of DialNav involves two models: Navigator and Guide. Navigator is a dialog-enabled agent that engages in conversation to gather additional navigation guidance, while Guide is a remote assistant with knowledge of the environment, responding to Navigator's inquiries. This section introduces the core capabilities of Navigator and Guide. Fig. 4 illustrates the overall process of DialNav, highlighting the interaction between Navigator and Guide. To study these capabilities, we modularize each component and leverage existing models for different functionalities.

4.1. Navigator

Navigator begins the task with an initial instruction I and, at each timestep, either updates $\mathcal{T}_t$ by taking a navigation action or updates $\mathcal{D}_i$ by posing a question. To effectively complete the task, Navigator requires modules, each handling one of the three core capabilities: (1) determining navigation actions, (2) identifying appropriate moments to ask questions, and (3) generating relevant questions.

Navigation At each navigation step t, Navigator selects the next node based on past dialog and navigation history. This is analogous to Vision-and-Language Navigation (VLN) [1], where an agent navigates to a goal based on a natural language instruction. Unlike VLN, where instructions are static and provided at the beginning, dialog in DialNav is dynamically collected across multiple visited nodes. We train a navigation module using a VLN model architecture, addressing this discrepancy by treating the past dialog as a single instruction for the remaining navigation path similarly to [31]. Specifically, we evaluate HAMT [3] and DUET [5], including pretrained weights from ScaleVLN [37]. DUET with pretrained weights, which showed optimal performance, is adopted as our baselines. Further details on model selection for each module are provided in Sec. L in Supp. Mat.

Whether to Ask Asking questions at appropriate moments is another capability for Navigator; frequent questions overwhelm Guide, while too few can lead to navigation errors. Striking a balance between navigation performance and dialog efficiency is a crucial mission in dialog-based navigation. In this work, we tested 3 approaches; Fixed-Interval [28], where questions are asked at regular navigation step intervals, Confidence Thresholding [13, 42], where questions are triggered when action confidence falls below a threshold, and Decision Head [29], which learns questioning timing from RAIN using action decision features. The baseline model adopts Decision Head.

Question Generation When uncertain about the next navigation step, Navigator asks Guide for additional information, often describing its surroundings to provide context. For question generation, we use LANA [36], a VLN instruction generation model that, like this task, generates descriptions based on visual landmarks. Additionally, we test LLaVA-1.5 [20], a multimodal LLM, with prompts for comparison. Although the target tasks differ, we pretrained LANA following [36], to generate navigation instructions. LANA serves as our baseline module for question generation.

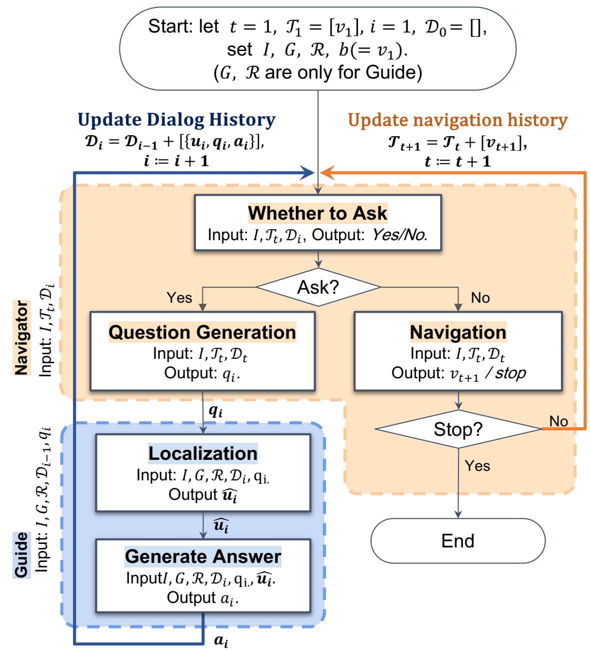

Figure 4. **Sequential interaction between Navigator and Guide.** This diagram illustrates the collaborative execution of DialNav. At each iteration, Navigator determines whether to ask a question. If a question is posed, Guide first localizes Navigator and then formulates a response. Each question-answer exchange updates the dialog history. If no question is asked, Navigator proceeds with navigation. The task concludes when Navigator decides to stop. $\hat{u_{i+1}}$ represents the viewpoint where Guide estimates Navigator to be based on localization.

4.2. Guide

Guide provides guidance to Navigator based on the environment graph G, initial instruction I, the goal region $\mathcal{R}$, and dialog history $\mathcal{D}_i$, which updates at each dialog turn i. Note that since Guide is remote, it does not have access to Navigator's navigation history $\mathcal{T}_t$.

Localization When Navigator poses a question, Guide must first estimate Navigator's location. This is similar to the Localization from Embodied Dialog (LED) task [15], which involves determining an observer's location based on dialog. We evaluate two ranking models from [15]: Simple Cross-Modal Network (SCN) and Graph Convolutional Network (GCN). After testing both models pretrained on the WAY dataset [15] and finetuned on RAIN, we adopt GCN as our baseline module.

Answer Generation Answer generation involves formulating responses that guide Navigator toward the destination. This primarily includes describing the next path to the goal region, similar to a VLN instruction generation. We evaluate LANA [36] and Llama-3.1-8B-instruct [32] with prompting for answer generation. Like in question generation, we adopt LANA as our baseline module, initializing it with pretrained weights following [36] and finetuning it on RAIN.

Setup	SR↑	OSR↑	SPL↑	NE↓	NSC	DTC	LE↓
Val Seen							
(1) Nav. only	18.2	28.6	17.3	14.5	19.6	-	-
(2) +Dialog	27.0	34.5	25.4	11.5	16.4	1.9	22.6
(3) +GT Loc.	**31.4**	**43.3**	**28.6**	**9.9**	18.3	1.7	-
Val Unseen							
(4) Nav. only	15.4	**33.3**	10.3	14.9	20.8	-	-
(5) +Dialog	13.9	20.7	10.0	16.3	14.2	2.9	21.9
(6) +GT Loc.	**19.8**	29.0	**15.5**	**12.7**	16.4	2.6	-
Test							
(7) Nav. only	12.7	**28.6**	8.0	14.8	22.8	-	-
(8) +Dialog	11.9	19.0	8.2	17.3	15.4	2.9	22.9
(9) +GT Loc.	**17.3**	23.7	**12.2**	**14.2**	17.4	2.7	-

Table 3. **Performance across dialog setup variations.** In **Nav. only**, the agent [5] navigates solely based on the initial instruction I, mirroring the standard VLN setting and its performance in DialNav. **+Dialog** enables dialog during navigation, corresponding to DialNav setup. **+GT Loc.** provides Guide with Navigator's ground-truth location, removing localization error.

5. Experiments

5.1. Experimental Settings

Training Modules Every module in Navigator and Guide is trained on RAIN. During training, each episode is processed for its respective subtask (*e.g.*, each QA turn forms a training example for question generation). We initialize each module with pretrained weights from related VLN tasks. Further training details are provided in Sec. L in Supp. Mat.

Cooperative Evaluation As depicted in Fig. 4, DialNav requires continuous collaboration between Navigator and Guide. Previously discussed modules are integrated in holistic dialog-based navigation. We employ pretrained DUET [5] for navigation, LANA [36] for question and answer generation and GCN for localization. All results reported with the average score of five runs.

Metrics We report four navigation metrics adopted from prior VLN works [1, 14, 16, 37]. **Success Rate (SR)** measures the percentage of episodes where Navigator successfully stops in the target region. **Oracle Success Rate (OSR)** counts episodes as successful if Navigator passes through the target region, benefiting from more exploration. **Success weighted by Path Length (SPL)** evaluates the success rate while considering path efficiency, penalizing unnecessary detours. **Navigation Error (NE)** measures the distance between Navigator's final position and the goal, indicating how close Navigator ends up to the goal region. Additionally, we report metrics for navigation and dialog efficiency, as well as localization accuracy. **Navigation Step Count (NSC)** is the number of steps taken by the model, thereby measuring the extent of navigation exploration. **Dialog Turn Count (DTC)** indicates the number of dialog turns in a single episode, reflecting the level of dialog engagement and efficiency. **Localization Error (LE)**, measures the distance between Guide's predicted location and Navigator's ground-truth position at each dialog turn, capturing cumulative errors from both question generation and localization.

Non-pretrained module	SR↑	NE↓	NSC	DTC	LE↓
Fully pretrained	27.0	11.4	16.4	1.9	22.6
−Navigation	9.2	14.4	20.8	1.7	22.0
−Question generation	27.7	11.0	16.4	1.9	20.8
−Localization	25.7	12.8	16.8	2.0	25.1
−Answer generation	11.7	14.8	16.9	1.9	23.5

Table 4. **Effects of pretrained modules.** Performance comparison when pretraining is omitted from specific modules. The 'Non-pretrained' column indicates the module without pretrained initialization, while 'Fully Pretrained' refers to the setup where all modules are pretrained.

5.2. Results

Enabling Dialog Tab. 3 compares models with and without dialog, highlighting its impact on visual navigation. In seen environments (Val Seen), the model without dialog (Row 1) relies solely on the ambiguous initial instruction, leading to poor navigation performance. In contrast, incorporating dialog (Row 2) enables active engagement, as reflected in DTC (1.9 dialog turns per episode), facilitating disambiguation and informed decision-making. As a result, navigation performance improves significantly, with relative gains of 8.8%, 5.9%, 8.2% and 3m in SR, OSR, SPL and NE respectively. We further tested a setup where the GT location of Navigator is provided to Guide when a question is posed. This simulates an omniscient guide, removing the need for well-formed questions and precise localization, similar to prior works [31]. The results show improved dialog efficiency with a lower DTC, as expected, and enhanced navigation performance due to more accurate answers enabled by perfect localization. These findings underscore the importance of well-posed questions and accurate localization in DialNav.

Generalization to Unseen Environments The trends observed in unseen environments (Val Unseen and Test) differ from those in seen environments. Overall, performance is poor in both settings, with and without dialog. We conjecture that this is due to the limited dataset scale, constrained by the high cost of data collection process compared to the complexity of DialNav. Addressing this challenge remains a key direction for future research. While both setups perform poorly, they exhibit different behaviors. The model without dialog (Row 4 and 7) explores the environment excessively, resulting in an unnecessarily large number of navigation steps (NSC). This leads to high OSR (favoring exploration) but low SR and SPL, as the task inherently lacks sufficient information to determine the correct stopping location.

In contrast, the model with dialog (Row 5 and 8) continuously seeks additional information, as it remains uncertain about the unseen environment. This behavior significantly reduces unnecessary exploration. However, SR remains low

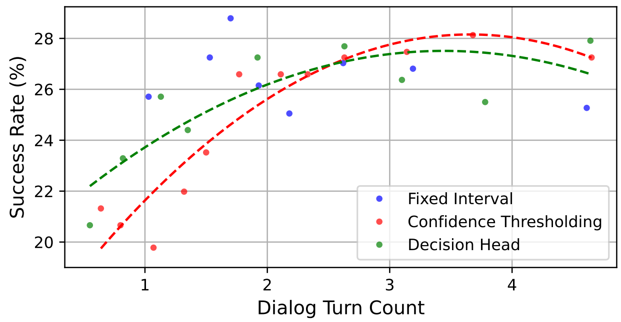

Figure 5. **Success Rate vs. Dialog Turn Count across three WTA strategies.** We evaluate three whether-to-ask (WTA) variants; Fixed Interval (blue), Confidence Thresholding (red), and Decision Head (green), each tested with varying the interval, confidence threshold, or logit scaling, respectively. The red and green dotted lines is fitted curves illustrating the trend of Confidence Thresholding (red), and Decision Head (green).

when the question and answer generation modules fail to produce grounded dialog. When the GT location of Navigator is given, navigation performance improves as the model no longer depends on question generation and localization. By removing the reliance on these components, error propagation is reduced, making performance primarily dependent on the answer generation and navigation modules, which benefit from strong pretrained modules. The impact of these pretrained modules is further elaborated below.

Effects of Pretraining To evaluate the impact of pretraining on each module, we conduct an ablation study by removing pretrained weights from individual components. Since each module is pretrained on a related but distinct task, its contribution varies depending on the similarity between the pretraining and target tasks, as well as the dataset scale. The results, presented in Tab. 4, demonstrate the critical role of pretraining in model performance. When the navigation module is randomly initialized (Row 2), SR and SPL drop significantly compared to Row 1, where pretrained weights are used for all components. Additionally, both navigation and dialog efficiency decline, leading to an increased NSC and DTC. Similarly, when the answer generation module is trained without pretrained weights (Row 5), navigation performance suffers despite a comparable NSC and DTC. This degradation is primarily due to the frequent generation of inaccurate guidance, causing misdirections for Navigator. These findings highlight the effectiveness of pretraining on relevant tasks, specifically instruction-following VLN tasks for navigation and instruction generation based on a given path for answer generation in DialNav. In the case of localization, the WAY [15] dataset is relatively small, making it insufficient to provide meaningful improvements. Pretraining the question generation also had a limited impact, likely due to misalignment between the pretraining task (instruction generation from paths) and the target task.

Question	Answer	SR↑	NE↓	LE↓	QF↑	AF↑
LANA [36]	LANA [36]	27.0	**11.4**	**22.6**	4.7	6.8
LLaVA [20]	LANA [36]	**27.3**	12.0	23.5	**9.0**	6.7
LANA [36]	Llama [32]	22.2	13.3	22.8	4.6	**9.0**

Table 5. **Performance across different question and answer generation modules.** LANA for question or answer generation is replaced with LLaVA-1.5 [20] or Llama-3.1 [32] with captions, respectively. **QF** and **AF** denote the fluency of questions and answers. LANA outperforms in NE and LE due to fine-tuning, while LLaVA and Llama generate more fluent QA owing to large-scale language training.

Impact of Whether-to-Ask Strategies We evaluate the impact of three whether-to-ask (WTA) variants by analyzing SR and DTC in Fig. 5: Fixed-Interval (blue dots), Confidence Thresholding (red dots) and Decision Head (green dots). While the Fixed Interval strategy exhibits a weak correlation between DTC and SR, both Confidence Thresholding and Decision Head show a positive correlation, enabling a trade-off between task success and dialog efficiency—an important consideration in human-in-the-loop scenarios, where excessive interaction may be perceived as intrusive. However, for both strategies, the benefits plateau after a few exchanges—beyond three turns for Decision Head and four for Confidence Thresholding—likely due to the modular nature of current dialog systems, which hinders context-aware interaction, and the navigation model's limited capacity to interpret and utilize dialog content effectively.

Dialog Fluency We also assess dialog fluency, by comparing LANA, fine-tuned for the generation of questions and answers, with LLaVA-1.5 for question generation and Llama-3.1-8B for answer generation, where the Llama-3.1-8B is prompted with a set of generated captions from panoramic images along the path (Tab. 5). Fluency scores for question and answer (QF and AF) are obtained by prompting Llama-3.1-8B on a 10-point scale (details in Sec. Q in Supp. Mat.). LANA scores below 5 due to limited dialog training data, producing repetitive and unstructured sentences. In contrast, LLaVA and Llama-3.1, trained on large-scale language data, exhibit higher fluency. LANA outperforms LLaVA and Llama in NE and LE by generating task-specific QA. A key future direction is integrating LLaVA/Llama communication capabilities into navigation tasks for task-specific, environment-grounded dialog.

Qualitative Results Fig. 6 presents a qualitative result of DialNav. Navigator asks questions to clarify the path and goal location. The generated questions incorporate key objects from the scene to aid in localization. Guide responds with relevant visual concepts, such as room names or visible objects, either along the path or near the destination allowing successful navigation.

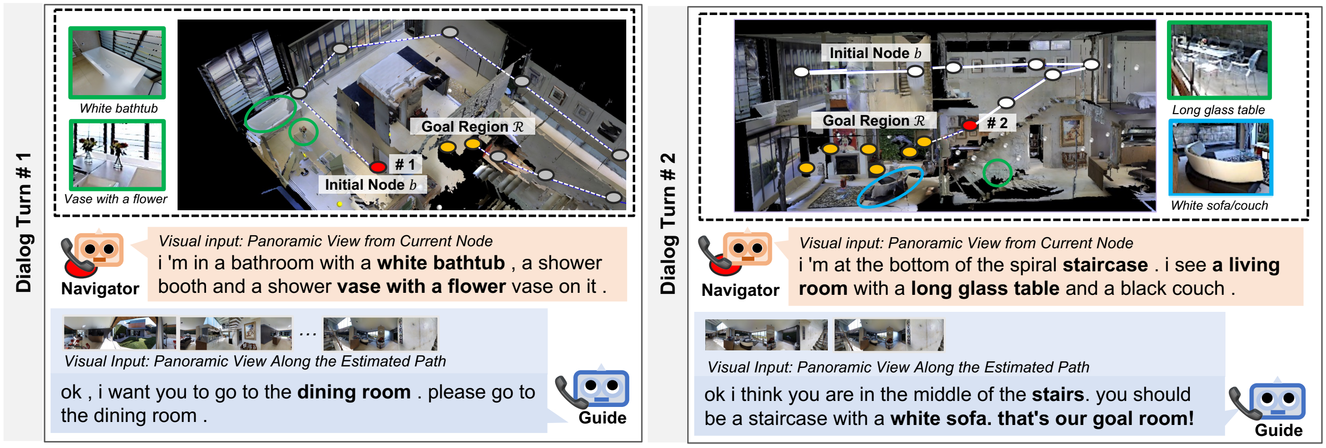

Figure 6. **A qualitative DialNav episode with the initial instruction: '*The goal room contains a bookcase*.'** The 3D reconstructed map at the top of each turn shows the same Matterport 3D environment from different angles, with ellipses representing nodes in Navigator's full path. A red node indicates where Navigator posed a question at the current turn, yellow nodes mark the goal region, and gray and white nodes represent unvisited and visited nodes, respectively, up to the current turn. Navigator formulates questions (orange boxes) with detailed descriptions of surrounding objects (highlighted with green circles in the 3D maps and corresponding images on the side), such as '*white bathtub*' in the first turn, or '*long glass table*' in the second turn. Guide provides visual hints (blue circles), like '*white sofa*' in the second turn, along with path guidance (blue boxes) and instructions on where to stop (*e.g.*, '*that's our goal room*'), which are crucial for achieving high SR. Through this dialog, Navigator successfully reaches the goal region.

6. Challenges

DialNav poses multiple challenges, making it particularly complex to tackle. Below, we elaborate on these challenges.

Expensive Data Collection The interdependency among subtasks requires holistic data collection with real-time interaction between two expert annotators, making it costly and difficult to scale. Additionally, the turn-taking nature of the task forces one annotator to wait for the other, resulting in an inefficient and time-consuming process. Furthermore, the limited diversity within the house simulation environment restricts scalability. The insufficient size and diversity of the dataset result in poor generalizability in unseen environments, particularly given the task's complexity. Future research should investigate automatic data creation and augmentation techniques for enhanced robustness.

Interdependancy between Subtasks The interdependent subtasks complicates submodule training, as even well-performing modules in isolation may fail within the full system due to error propagation. Minor errors in one module can cascade through subsequent stages, leading to inaccurate navigation decisions. Thus, mitigating error propagation and ensuring robust performance across all modules is crucial for successful task execution. Training the system as a whole could potentially mitigate these issues, but developing a unified, end-to-end model is also a significant challenge.

Evaluation in Dynamic Context Evaluating DialNav presents significant challenges due to its dynamic and non-deterministic nature. Both dialog generation and navigation involve sequential, non-definitive predictions, where slight contextual differences can significantly change the future outputs. As a result, the collected GT annotations may become invalid or unavailable for evaluation in a predicted context, further complicating system assessment. Addressing these challenges requires flexible, context-aware evaluation methods that can adapt to the evolving nature of the task.

Long-form Multimodal Context Modeling DialNav requires multimodal sequential modeling, processing panoramic images, action trajectories, and multi-turn dialog. This complex dependency significantly increases task difficulty, making it challenging to integrate and utilize information effectively over extended interactions.

Balance between Navigation and Dialog Effective navigation in DialNav requires striking a balance between task success and communication efficiency. Frequent interactions with Guide can improve the success rate, but excessive dialog may be inefficient or disruptive, while minimizing dialog may lead to increased navigation errors. Moreover, the agent must dynamically adapt its questioning strategy based on the reliability of Guide.

7. Conclusion

We introduced DialNav, a dialog-based navigation task highlighting the importance of conversation due to the guide's non-omniscient nature. To support this, we collected RAIN, a dataset of 2,231 human-human dialogs with navigation trajectories. We conducted experiments analyzing the current state of the art and discussed key challenges of DialNav. We will publicly release the dataset, code, data collection tool, and evaluation framework to foster future research.

References

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton 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. 2, 3, 5, 6

[2] Shurjo Banerjee, Jesse Thomason, and Jason Corso. The robotslang benchmark: Dialog-guided robot localization and navigation. In *Conference on Robot Learning*, pages 1384–1393. PMLR, 2021. 2

[3] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. *Advances in neural information processing systems*, 34:5834–5847, 2021. 5, 2

[4] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Learning from unlabeled 3d environments for vision-and-language navigation. In *ECCV*, 2022. 2

[5] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. 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*, pages 16537–16547, 2022. 5, 6, 2

[6] Ta-Chung Chi, Mihail Eric, Seokhwan Kim, Minmin Shen, and Dilek Hakkani-tur. Just ask:an interactive learning framework for vision and language navigation, 2019. 3

[7] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–10, 2018. 2

[8] Davison. Real-time simultaneous localisation and mapping with a single camera. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1403–1410. IEEE, 2003. 2

[9] Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. Talk the walk: Navigating new york city through grounded dialogue, 2018. 2

[10] Yue Fan, Winson Chen, Tongzhou Jiang, Chun Zhou, Yi Zhang, and Xin Eric Wang. Aerial vision-and-dialog navigation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3043–3061, Toronto, Canada, 2023. Association for Computational Linguistics. 2

[11] Yue Fan, Jing Gu, Kaizhi Zheng, and Xin Eric Wang. R2h: Building multimodal navigation helpers that respond to help requests, 2023. 2

[12] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. *Advances in neural information processing systems*, 31, 2018. 2

[13] Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. Dialfred: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters*, 7(4):10049–10056, 2022. 2, 3, 5

[14] Meera Hahn and James M Rehg. Transformer-based localization from embodied dialog with large-scale pre-training. *arXiv preprint arXiv:2210.04864*, 2022. 2, 6

[15] Meera Hahn, Jacob Krantz, Dhruv Batra, Devi Parikh, James M Rehg, Stefan Lee, and Peter Anderson. Where are you? localization from embodied dialog. *arXiv preprint arXiv:2011.08277*, 2020. 2, 5, 7, 6

[16] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. 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*, pages 13137–13146, 2020. 6

[17] Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. Stay on the path: Instruction fidelity in vision-and-language navigation. *arXiv preprint arXiv:1905.12255*, 2019. 2

[18] Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. *arXiv preprint arXiv:2010.07954*, 2020. 2

[19] Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. Topviewrs: Vision-language models as top-view spatial reasoners. *arXiv preprint arXiv:2406.02537*, 2024. 2

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 5, 7, 6

[21] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16488–16498, 2024. 2

[22] Anjali Narayan-Chen, Prashant Jayannavar, and J. Hockenmaier. Collaborative dialogue in minecraft. In *Annual Meeting of the Association for Computational Linguistics*, 2019. 2

[23] Khanh Nguyen and Hal Daumé III. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. *arXiv preprint arXiv:1909.01871*, 2019. 2, 3

[24] Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12527–12537, 2019.

[25] Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2017–2025, 2022. 2, 3

[26] Seth Pate and Lawson LS Wong. "the wallpaper is ugly": Indoor localization using vision and language. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1558–1564. IEEE, 2023. 2

[27] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020. 2

[28] Homero Roman Roman, Yonatan Bisk, Jesse Thomason, Asli Celikyilmaz, and Jianfeng Gao. Rmm: A recursive mental model for dialog navigation. *arXiv preprint arXiv:2005.00728*, 2020. 2, 3, 5

[29] Kunal Pratap Singh, Luca Weihs, Alvaro Herrasti, Jonghyun Choi, Aniruddha Kembhavi, and Roozbeh Mottaghi. Ask4help: Learning to leverage an expert for embodied tasks. *Advances in Neural Information Processing Systems*, 35:16221–16232, 2022. 2, 3, 5

[30] Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. Executing instructions in situated collaborative interactions. In *Conference on Empirical Methods in Natural Language Processing*, 2019. 2

[31] Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR, 2020. 2, 3, 4, 5, 6

[32] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 5, 7, 6

[33] Hanqing Wang, Wei Liang, Jianbing Shen, Luc Van Gool, and Wenguan Wang. Counterfactual cycle-consistent learning for instruction following and generation in vision-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15471–15481, 2022. 2

[34] Haoyu Wang, Le Wang, Sanping Zhou, Jingyi Tian, Zheng Qin, Yabing Wang, Gang Hua, and Wei Tang. Towards precise embodied dialogue localization via causality guided diffusion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13350–13360, 2025. 2

[35] Su Wang, Ceslee Montgomery, Jordi Orbay, Vighnesh Birodkar, Aleksandra Faust, Izzeddin Gur, Natasha Jaques, Austin Waters, Jason Baldridge, and Peter Anderson. Less is more: Generating grounded navigation instructions from landmarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15428–15438, 2022. 2

[36] Xiaohan Wang, Wenguan Wang, Jiayi Shao, and Yi Yang. Lana: A language-capable navigator for instruction following and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19048–19058, 2023. 2, 5, 6, 7

[37] Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Hao Tan, and Yu Qiao. Scaling data generation in vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12009–12020, 2023. 5, 6, 2

[38] Yu Yan, Rongtao Xu, Jiazhao Zhang, Peiyang Li, Xiaodan Liang, and Jianqin Yin. Instrugen: Automatic instruction generation for vision-and-language navigation via large multimodal models, 2024. 2

[39] Chao Zhang, Mohan Li, Ignas Budvytis, and Stephan Liwicki. Dialoc: An iterative approach to embodied dialog localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12585–12593, 2024. 2

[40] Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, and Xiaodan Liang. Soon: Scenario oriented object navigation with graph-based exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12689–12699, 2021. 2

[41] Wanrong Zhu, Xin Eric Wang, Tsu-Jui Fu, An Yan, Pradyumna Narayana, Kazoo Sone, Sugato Basu, and William Yang Wang. Multimodal text style transfer for outdoor vision-and-language navigation. *arXiv preprint arXiv:2007.00229*, 2020. 2

[42] Yi Zhu, Yue Weng, Fengda Zhu, Xiaodan Liang, Qixiang Ye, Yutong Lu, and Jianbin Jiao. Self-motivated communication agent for real-world vision-dialog navigation. in 2021 ieee. In *CVF International Conference on Computer Vision (ICCV). IEEE, Montreal, QC, Canada*, pages 1574–1583, 2021. 2, 5

DialNav: Multi-turn Dialog Navigation with a Remote Guide

Supplementary Material

H. Data Collection Interface

Fig. G illustrates the data collection interface for Navigator and Guide.

Navigator Interface Navigator moves toward the goal region given the initial instruction. When unsure, Navigator can ask questions to receive additional information from the remote Guide. When Navigator asks a question, Navigator's interface is deactivated until Guide's response comes. Navigator interface consists of four main components:

- **Navigation Interface**: Allows free movement between nodes, active only during the Navigator's turn.
- **Hint**: Provides hints about objects in the goal room, formatted as *"The goal room contains ⟨object⟩."*
- **Chat Interface**: Enables the Navigator to send questions

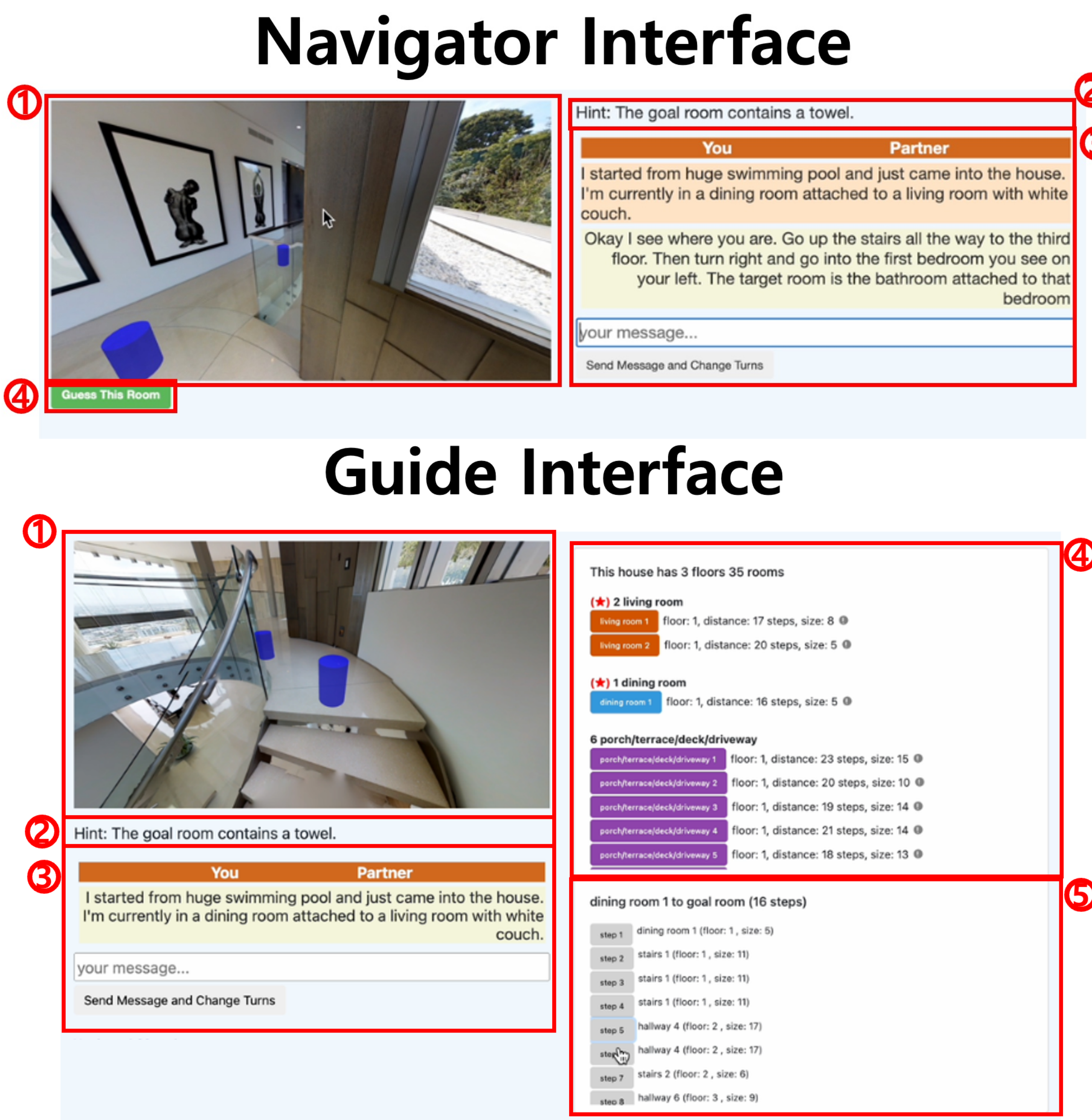

Figure G. Navigator (top) and Guide (bottom) interface

and receive responses from the Guide.
- **Guess Button**: Allows the Navigator to indicate they have reached the goal. If correct, the task ends; if incorrect, a popup indicates the mistake, and the task continues. This button can be pressed multiple times without changing turns.

Guide Interface Guide acts as a knowledgeable participant regarding the house environment. We designed a specialized Guide interface that simulates Guide to be familiar with the entire space. For the question of whether Guide interface provided sufficient support for the task, the annotators rated 4.5 out of 5 on average. Guide interface consists of five components:
- **Navigation Interface**: Allows free movement between nodes, even when it is not Guide's turn.
- **Hint**: The same hint provided to Navigator is also shared with Guide.
- **Chat Interface**: Enables Guide to read the Navigator's questions and send responses.
- **House Info Interface**: Offers detailed information about the house, including:
  - **Overall Information**: Presented as *"This house has N floors and N rooms."*
  - **Room List**: Displays information about all rooms grouped by type, including size, floor location, included objects, and distance to the goal region. Clicking on a room name allows Guide to view and navigate nearby areas.
  - **Shortest Path Trajectory**: The shortest path from the current location to the goal region, detailing nodes, rooms, and floors along the path. Clicking a node navigates Guide to that node in the navigation interface.

I. Data Collection Details

We recruited annotators from volunteers within the university, though participation was not limited to students. All participants were between the ages of 20 to 35. Each participant received a compensation of 50,000 KRW for participating in the experiment for a continuous 3.5-hour session, which is equivalent to 1.45 times the minimum hourly wage.

J. Data Preprocessing

A total of 2,421 episodes were conducted. To ensure dataset quality, we first filtered out episodes that contained technical faults or exceeded the predefined time limit of 22 minutes. Additionally, we manually reviewed outlier episodes based on three criteria: trajectory detour, dialog count, and annotator scores. Specifically, we identified episodes within the top 1% of outliers for each metric—5.6 for detour, 6 for dialog count, and 1 for annotator scores. Upon review, these episodes were retained, as they were deemed sufficiently valid for the task. Fig. H and Fig. I presents examples of qualitative DialNav samples and outliers respectively. To further refine the dataset, we manually corrected instances where dialog entries were inadvertently split due to user errors, such as accidental key presses. After these refinements, the final dataset consisted of 2,231 episodes. The dataset was then partitioned into training, validation, and test splits, with 1,559 episodes for training, 285 for testing, and 387 for validation. The validation set was further divided into seen (111 episodes) and unseen (276 episodes) environments, while the test set exclusively contained unseen environments. We adhered to the seen/unseen/test environment split established in prior work [1].

K. The RAIN-Segment Dataset

A RAIN episode consists of multiple dialog turns and corresponding actions. For training, we restructured episodes into segment instances following the methodology of CVDN [7]. Specifically, an episode with three dialog turns generates four segment instances: one representing the trajectory before any dialog and three corresponding to successive dialog turns.

Each RAIN-Segment is defined as $(G, I, \mathcal{R}, \mathcal{T}_t, \mathcal{D}_i)$, where G, I, and $\mathcal{R}$ are common across the entire episode, while $\mathcal{D}_i$ is dialog history until dialog turn i and $\mathcal{T}_t$ is trajectory history up to dialog turn i. Each RAIN-Segment instance represents a specific state within an episode, capturing the interactions between the Navigator and the Guide.

For Navigation training, all instances are utilized, while for Localization and QA generation, first segment in each episode which does not contain dialog is not included. We adhered to the seen/unseen/test environment split established in prior work [1]. Refer to Tab. F for the respective counts.

Type	Train	ValSeen	ValUnseen	Test	Total
DialNav	1,559	111	276	285	2,231
Instance	4,493	337	805	768	6,403
w/ dialog	2,934	226	529	483	4,172

Table F. Dataset size per splits

L. Module-wise Training and Experiments

Navigation. We utilized two existing VLN models for navigation: HAMT [3] and DUET [5]. HAMT is first initialized with publicly available pretrained weights, and finetuned on DialNav training instances. To evaluate the impact of the large-scale pretraining on DialNav navigation, we compared the DUET with weights pretrained on ScaleVLN [37]. For each RAIN-Segment instances, the initial instruction and the dialog history is appended to form a navigation instruction for each instance. The VLN models were trained to determine the best next optimal action given the instruction. Due

Q: Hi I'm in a grey bedroom. The bed has dark brown frame and grey patterned sheets. There's a grey sofa at the foot of the bed, facing a TV. I can see two big windows.

A: Please exit the room using the door on the right. Then walk down the hallway and you'll find a kitchen on your left and a living room with sofas on your right. Go through them and you'll see pictures of a couple and their child on the wall. There are 6 in total with different shapes. Do you see them?

Q: Yes. I'm standing in front of them.

A: Ok now take a right turn and walk a few steps ahead. Then when you see two pictures of a boy and a girl on the wall, make a right turn and you'll see a bedroom. Our destination is the bathroom attached to the bedroom

Q: I was standing in the corner, which has two ways to go. There are lots of rooms and they look all same. I entered a nearest room, and I can see a white table and a black chair. On the door, number '321' is written.

A: Great. Go out of the room and turn left. Walk straight down the hallway up until you see 4 red sofas on your left. Tell me when you get there

Q: I can see the red sofas. I am facing them.

A: Okay. Turn right there. Having the red sofas on you back, walk down the hallway up until you see a white pillar and a poster of a men attached to the glass wall on you left. And a exit sign on your right

Q: I am there! Facing the exit sign.

A: Great! Here's the tricky part. Walk more down the hallway, until you find a messy room with a mircrowave on your right, and in front of the goal room, there's another small office with only one white desk and chair.

Q: Hello, I am in a small bathroom with a toilet and a black sink. There is a door to my right which is wooden brown, and also another door to my left which is white.

A: Alright, go through the white door, and go straight at the sofa. If you stand in front of a table with fruits, that's the goal room.

Q: now im in a bedroom; a bed with patterned bedding, an abstract painting of two horses right above the bed, a black chair set in front of the bed.

A: You're on the 4th floor and we're gonna have to make a way to the 2nd floor. So I want you to come out of the bedroom through the door on your right: to the TV side. Once you're out, there will be a lounge with a round wooden table and a staircase going down on your 2 o'clock. We'll take down those round stairs and let me know when you're down and find another stairs in front of you.

Q: yep now im facing another stairs

A: Yes head down the stairs to the entry lobby and make a U-turn to your left and there will be another short stairs for you to go down. Once you're down again from those stairs, go to the living room that you can find on your 2 o'clock: right side of the statue at your front. Once you've made your way to the living room, take a right turn and head to the pool table. There's a bathroom on the right side of that pool table and that should be our goal room.

Q: I'm in a very fancy living room with a greyish blue fireplace with a mirror above it. There is also a chandelier, and the pattern on the floor has some red in it. Should I leave the room into a long hallway?

A: There are several rooms that have a greyish blue fireplace, so could you please elaborate on it more? For example, it has a brown wooden round table, it has a white striped patterned sofas, like that. Thanks.

Q: Oh, sorry. It has a long white couch, with 5 small white chairs aligned next to it. There is also a small bed in the corner.

A: Oh, I found you. Enter the hallway, which is on the left side of the fireplace. If you walk through the hallway, you'll see the stairs on your right side. Could you go halfway down the stairs?

Q: Yes, I'm halfway down the stairs. Keep going?

A: You have to stop. Is the stair you've walked down on your behind?

Q: Yes, and there are stairs on either side behind me. I'm facing the white wall.

A: Great. Now just take one step at the stair in right side of you. That is the goal room.

Figure H. **Sample RAIN dialogs**

Q:	I'm in a room with a swimming pool. There are 3 sunbeds, and I see a treadmill on one side.
A:	good new. it's near. assuming you are facing the swimming pool, turn right and take a few step and turn right and you will see two spa rooms: take the one on the left and the room has this fan with red flowers and a blanket with check patterns. Text me if you aren't confused!
Q:	I'm looking at the swimming pool, and there are sunbeds behind me. If I turn right and another right, I see a concierge, not spa rooms.
A:	ok. there are two pools right? Stand in front of the larger pool. (The one on the left) and you see that there are plants next to that pool. There are two rooms behind those plants. I need you to go behind that plant, so I need you to head towards the diamond shape wall. From here, turn right. Let me know if you find the diamond wall.

Q:	Hi. I'm in the room with a huge bed which has a purple bedding, and I can also see the light green colored sofa. I guess I'm at the second floor.
A:	Hi. Is the roof inclined not flat? And do you see trees through a window?
Q:	Yes. The roof is inclined, and I can see tress through the window.
A:	Great! You're currently on the third floor. First come out from the bedroom. While you come out from that room, trees should be on your left side.
Q:	Okay. Should I go down stairs?
A:	Yes. Please go down the stairs.
Q:	I came down. Now I'm standing at the kitchen, and can see a big wood table.
A:	Terrific! You will see a hallway at the left side of that big wood table. Please go to the hallway.
Q:	I'm at the hallway now. I can see a pool through the window.
A:	Sorrry! Do you see a black car through a window?
Q:	I guess I was at the different hallway. I got to the other one, and now I can see the black car through the window.
A:	Fantastic! There are brown stairs between white walls. Please go down to the first floor by following those stairs.
Q:	I came down! Now I'm looking ahead a room with work out machines.
A:	Good. Please do not go into that room. Instead enter a room next to the room with workout machines. Then you will see a wall on which a kind of graffiti is drawn.
Q:	I'm in!
A:	The place in front of that wall. That is the goal.

Q:	in front of stairs going down, on my left on the wall is an abstract painting, signiture reads what looks like 'Hunk'
A:	i got it. it's quite complicated path to our room. putting that painting on your right, please make a right turn.(kind of u turn you should make) then you might see set of wooden stairs leading to upstairs. going up, you can make a right turn to see a few black stairs. finishing that stair, there might be a kitchen. please make a right turn right at the person like-statue. then there might be a set of wooden stairs, and please go upstairs.finisnihing that, you might find a bed room and going inside, there is a spa. that's our goal room.
Q:	there's no right turn I can make if I put the painting on my right. Im in front of stairs going down, behind me is an open white door (a room with two big yellow sofa beds)
A:	oh isn't that the yellow painting?
Q:	yes, if i make a right turn, i'm in a sauna. just tell me which floor I need to be on. do i need to go downstairs?
A:	our goal room is on the 3rd floor. if that's wooden spa, i guess its our goal room

Q:	I'm at front of stairs, do I have to go down?
A:	There are 3 stairs in this house. I'll need more infos. Please give me more details.
Q:	I can see two bedrooms, one has grey blanket on the bed, and the other one has brown blanket on the bed.
A:	So, you're in hallway 1, floor2. Our goal room is bedroom 4. Please enter bedroom with the brown blanket. That's our goal room.

Figure I. **Sample dialog with outliers**: (1) Dialog with a significant detour, (2) Dialog with a high number of QA turns, (3) Dialog with a low Guide score, (4) Dialog with a low Navigation score.

Method	+SV	Val Seen	Val Unseen	Test
Shortest		22.51	23.06	25.23
Random		1.91	1.94	0.25
HAMT [3]		11.26	8.31	5.87
DUET [5]		11.13	10.22	11.53
DUET [5]	✓	**12.88**	**12.07**	**12.37**

Table G. **Navigation scores on RAIN-Segment.** The goal progress of navigation to the destination is measured when the initial instruction and the last answer from the previous dialogue are given. **+SV**: Pretrained on ScaleVLN [37].

	Val Seen		Val Unseen		Test	
Method	**B4 ↑**	**RG ↑**	**B4 ↑**	**RG ↑**	**B4 ↑**	**RG ↑**
LLaVA [20]	.0311	.2074	.0318	.2183	.0259	.2043
LANA [36]	.0405	**.2087**	**.0533**	**.2203**	.0502	**.2131**
LANA(pt) [36]	**.0532**	.2005	.0527	.2111	**.0539**	.2071

Table H. **Question Generation scores on RAIN-Segment.** The similarity between the questions generated by the model and those generated by humans in the same context is measured. **B4**: BLEU4, **RG**: Rouge-L

	Language			Navigation	
Method	**B4 ↑**	**RG ↑**	**CD ↑**	**GP↑**	**SR↑**
Val Seen					
Llama-3.1	.0268	.1802	.0595	5.36	18.92
LANA	.0539	.2156	.0953	6.89	23.42
LANA(pt)	**.0745**	**.2342**	**.1310**	**10.63**	**30.09**
Val Unseen					
Llama-3.1	.0271	.1765	.0394	5.35	13.01
LANA	.0542	.2082	.0722	8.03	22.52
LANA(pt)	**.0596**	**.2223**	**.1086**	**8.92**	**25.28**
Test					
Llama-3.1	.0280	.1881	.0550	6.74	13.08
LANA	.0519	.2217	.0789	6.93	10.55
LANA(pt)	**.0648**	**.2325**	**.0952**	**9.50**	**17.11**

Table I. **Answer generation scores. LS:** Llama-S, **LF:** Llama-F, **GP (U):** GP on unseen validation set, **GP (S):** GP on seen validation set. The navigation scores (GP, SR) report the performance of the navigation task based on the answers generated by the model, using the DUET model pretrained with ScaleVLN.

to the multi-turn nature of the dialog, the answers often do not contain the complete path to the goal. Therefore, Goal Progress (GP) [31], the distance that agent got closer towards the goal, is used as navigation metric.

Tab. G shows the navigation performance on RAIN-Segment. We report shortest path length and random agent performance serving as an upper and lower bounds respectively. The DUET [5] model, which retains past trajectory history as a graph, demonstrated superior performance in unseen environemnt compared to the HAMT [3] model. Additionally, pretraining on ScaleVLN [37] further improved navigation performance across all environment splits. This result indicates the advanced model and large-scale pretraining on VLN models brings high performance in DialNav navigation as well.

Through these experiments, we selected DUET [5] pretrained with ScaleVLN [37] which showed the best performance across all environments for our baseline model.

Question Generation. We employed the VLN instruction generation model LANA [36] and the multimodal model LLAVA-1.5 [20] for generating questions. We tested LANA with and without pretraining . Although Navigator has access to the past trajectory and previous dialogs, to simplifying the task, we provided only the visual input from the current viewpoint in this work. LLAVA-1.5 (7B) is simply prompted to generate question including details of current panoramic view.

The objective of Question Generation is to facilitate human-like dialog; therefore, we evaluated the similarity of the generated questions to those asked by humans in identical situations. Tab. H, reports metrics for question generation. As discussed in the main manuscript, the improvement by pretraining LANA is limited in question generation due to the significant mismatch between pretext and target tasks.

Through these experiments, we selected LANA [36] with pretrained weight for our baseline model.

Whether to Ask. Whether to Ask (WTA) is a task to predict binary decision of to ask or not given navigation context. Previous dialog-based studies have used simple Fixed-Interval [28], Confidence Thresholding [13, 42] or adopting an additional model [29]. We tested all these 3 approaches. In Confidence Thresholding, questions are triggered when action confidence falls below a threshold. For the third approach, we added a decision head to DUET [5] model to utilize the intermediate output of the action decision and trained with RAIN-Segment. We selected the decision head for our baseline model.

Localization. Localization task predicts location of Navigator given the previous dialog. Although Guide has access to the full dialog history, we only use the last question in dialog for this work for simplification. We evaluate two ranking models from [15]: Simple Cross-Modal Network (SCN) and Graph Convolutional Network (GCN). Both compute node embeddings for the environment graph G and obtain a cross-modal feature by element-wise multiplying it with the

Method	Val Seen		Val Unseen		Test	
	LE↓	A@3↑	LE↓	A@3↑	LE↓	A@3↑
Random	19.76	7.96	18.77	4.34	21.54	3.46
SCN	**10.20**	**44.69**	14.63	23.58	16.04	19.14
SCN†	10.97	42.04	13.87	23.58	15.60	22.20
GCN	12.09	38.94	13.33	24.91	16.53	16.90
GCN†	11.47	36.73	**11.63**	**31.19**	**15.28**	**22.77**

Table J. **Localization scores on RAIN-Segment.** Scores are reported with the model with least LE for in Validation Unseen set in each model. **LE**: Localization error in meters. **A@3**: Accuracy with an allowable error margin of 3m.

query embedding, followed by a linear layer for ranking. The key difference is that SCN models each node independently, whereas GCN captures graph structure. All models are first pretrained on the WAY [15] dataset and then finetuned on RAIN-Segment. For evaluation, we adhere to the protocols in [14, 15], measuring localization error in meters and the accuracy of successful localization within 3 meters.

The SCN exhibits a significant performance disparity between seen and unseen environments, with minimal performance gains from WAY pretraining. (Tab. J ln2, 3) In contrast, the GCN model demonstrates greater gains from pretraining and better generalization to unseen environments. (Tab. J ln4, 5) This suggests that more data and complex models are required to generalize effectively to novel environments. The localization task remains underexplored, necessitating further research.

Through these experiments, we selected GCN with pretrained weight for our baseline model.

Answer Generation. We tested VLN instruction generation model LANA with and without pretraining. We also tested Llama-3.1-8B-instruct [32] with promppting to test power of large language model for answer generation task. For Llama, we first generates detailed caption of each navigation nodes with LLaVA-1.5 [20] and prompts Llama to generate navigation instruction based on the sequence of captions through route. See Fig. M and Fig. N for LLaVA and Llama prompts. Although Guide has full access to previous dialog and entire house map, we only use the remaining trajectory to the goal as input to simplified the task.

The objective of Question Generation is to facilitate human-like dialog; therefore, we evaluated the similarity of the generated answers in identical situations. Answer generation is similar to the VLN instruction generation task, so metrics improved through pretraining. To check the impact on navigation, we also reported goal progress and SR of DUET [5] model pretrained with ScaleVLN [37] when the answers were given as instructions.

Through these experiments, we selected LANA [36] with pretrained weight for our baseline model.

M. Question Generation Examples

Ground Truth (GT): I'm at a bedroom. There is a purple sofa. The bed has red blankets, one purple cushion and two white pillows. The outside is clearly showing, with trees and multiple cactus.

LLaVA: I'm in a room with a large window, and I can see a view of a desert landscape outside.

LANA: I'm in a hallway with a large glass window. Next to the sink is a glass shelf to the outside. There is a glass door to the right next to the sink. To the sink, there is a glass shelf to the kitchen.

Figure J. Qualitative analysis of question generation from different models.

Fig. J shows example of question generation given a viewpoint.

N. Answer Generation Examples

Fig. K shows an example of answer generation based on a list of panoramic views along the route to the goal room.

O. LLaVA Prompt for Question Generation

The prompt for LLaVA-1.5 used in question generation is shown in Fig. L. We provide LLaVA with a panoramic image to generate a scene description as a Navigator question.

P. Llama based Answer Generation

P.1. Generating LLaVA captions

The prompt for LLaVA-1.5 caption generation is shown in Fig. M. We provide LLaVA with a list of panoramic images and extract captions for the images corresponding to the path.

P.2. Prompting Llama to generate answers

The prompt for LLaMA-3.1-8B Instruct answer generation is shown in Fig. N. The scene description extracted from LLaVA, along with the corresponding prompt, is provided as input to Llama with an one-shot example.

Q. Llama Prompt for Language Evaluation

The prompt for LLaMA evaluation is shown in Fig. O. This prompt is used to evaluate fluency scores for both questions and answers (QF and AF).

Ground Truth (GT): Almost there. Now walk towards the light grey sofa in front of you. On the left to that sofa, there is a path to a terrace where you can see the pool on the left. The terrace is our target

LLaVA: Head towards the large window with the modern design, and you'll find the dining area with a table and chairs. The kitchen is adjacent to the dining area, and you can access it through a glass door. If you'd like to explore more, the staircase leading to another floor is located in the living room, which is decorated with a vase, a TV, and several books.

LANA: great . so , facing the white couch , you will see a grey couch on your right , and a tv on your right . go straight until you see a living room with a tv on your right . that 's our goal room .

Figure K. Qualitative analysis of answer generation from different models.

Prompt:
`You will be given a panoramic image of indoor scene. Create a sentence including what type of room the given image is (ex. bedroom, bathroom, empty room, stair, hallway), and details and any unique objects that would not appear in other regions of the building so that someone else can easily locate. Start with 'I'm in' or 'I can see'.`

LLaVA Output:
I'm in a room with a wooden door and a blue curtain.

Figure L. **Prompt for LLaVA's question generation and its results.**

Prompt:
`Please describe the region in this image.`

LLaVA Output:
The bathroom in the image is a small space with a toilet and a sink. The sink is located near the toilet, and there is a book placed on the countertop. The bathroom appears to be part of a larger living area, as it is situated next to a wall and a door.

Figure M. **Prompt for LLaVA's caption generation and its results.**

Prompt:
`You are an agent for creating navigation route. Given sequence of scene image description, create a navigation guide sentence for the route. You don't have to describe every single step. Try to add any unique object or landmark. Return your evaluation results in the following JSON format without any additional text:`
`'response': '<your response>'`

`[One-Shot Example]`

Llama Output:

Figure N. **Prompt for Llama's answer generation and its results.**

Prompt:
You will be given a sentence, and your task is to evaluate its fluency. Fluency refers to how natural and grammatically correct the sentence sounds in English. Rate the sentence on a scale from 1 to 10, where:
1: Very unnatural, severe grammar or structural errors.
10: Perfectly natural, indistinguishable from a native speaker's writing.

Your output should be in JSON format, containing your evaluation results based on the criteria above.

Figure O. **Prompt for Llama evaluation of fluency scores for question and answer(QF and AF).**