Advancing DialNav through Automatic Embodied Dialog Augmentation

Leekyeung Han[1] **Sangwon Jung**[2] **Hyunji Min**[1]
Jinseong Jeong[1] **Minyoung Kim**[1] **Paul Hongsuck Seo**[1]
[1]Korea University [2]Trillion Labs
{happilee12, daream2, dw9030, omniverse186, phseo}@korea.ac.kr
sangwon.jung@trillionlabs.co

Abstract

For embodied agents capable of physical interaction, the capability to create and understand dialog is crucial to ensure both safety and effectiveness. While DialNav (Han et al., 2025) provides a framework for holistic evaluation of the dialog–execution loop in photorealistic indoor navigation, its performance remains limited by a critical scarcity of training data (2K episodes). To address this, we propose an automatic generation pipeline, and construct the **RAINbow** dataset, a large-scale training dataset with 238K episodes for DialNav. Our pipeline converts existing VLN datasets into multi-turn dialog and creates cost-efficient and high-quality dataset. Then, we introduce two additional complementary advances to unlock the data's full potential: (1) Dual-Strategy Training, a navigation training scheme to align the navigation training with the dynamic dialog-navigation loop, and (2) a localization model that leverages VLN knowledge. By combining these complementary solutions, our model substantially outperforms the baseline in success rate on both **Val Seen** (58.24, **+89%**) and **Val Unseen** (29.05, **+100%**) splits, establishing a new state of the art. Our code and dataset are available at: https://happilee12.github.io/RAINbow

1 Introduction

Embodied agents must operate with high reliability, since misinterpreting instructions or executing incorrect actions can cause physical harm. Enabling dialog improves both safety and task effectiveness: by asking questions and refining their understanding before acting, agents can resolve ambiguity and adapt to dynamic situations. Learning dialog-enabled agents, however, remains highly challenging. Collecting datasets is costly, as it requires two people to coordinate in real time while grounding their conversations in the given task context. Furthermore, even with such data, training remains difficult because if either the action trajectory or the dialog deviates from the collected annotations, the supervision becomes invalid.

We study these challenges in **DialNav** (Han et al., 2025), a cooperative dialog-based vision-and-language navigation (VLN) (Anderson et al., 2018) task. VLN is an embodied navigation problem in which an agent follows a fixed natural language instruction given at the start of an episode to reach a goal. DialNav extends this into a dialog-based setup, where navigation unfolds through an interactive dialog exchange.

In DialNav, two agents, a *Navigator* and a remote *Guide*, collaborate through dialog to accomplish the navigation task in a photo-realistic indoor environment (Matterport3D (Chang et al., 2017)). The Navigator starts with only an ambiguous and underspecified instruction (*e.g.,* "the target room has a plant"), which is insufficient to reach the goal; success therefore hinges on the following dialog. To this end, the Navigator proactively asks the Guide for additional guidance during navigation whenever uncertainty arises. Due to the remote Guide setup in DialNav, the Guide first has to *localize* the Navigator based on this question, which makes the question critical. The Guide then answers with context-specific instructions that help subsequent navigation decisions. To train such agents, Han et al. (2025) first collected the RAIN dataset, a human-human interaction dataset for DialNav. Then, the Navigator and Guide are modeled by separated neural modules specialized for their constituent capabilities (*e.g.,* question generation, answer generation, and navigation). These modules are trained separately on RAIN and later integrated for *holistic* navigation-dialog loop of interactive navigation.

While Han et al. (2025) introduced the first dataset and baseline models for DialNav, their performance remains limited. A key reason is the severe scarcity of training data: RAIN contains only 2K episodes, which is insufficient to support

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 44876–44904
October 24-29, 2026 ©2026 Association for Computational Linguistics

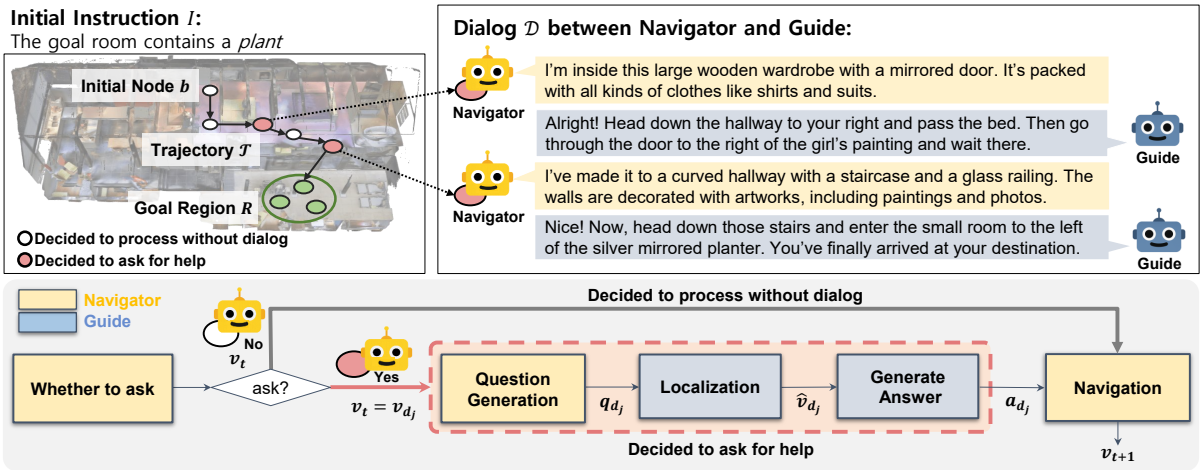


Figure 1: **Overview of the DialNav task**. **Top:** The Navigator starts at an initial node b and navigates to reach the goal region R. Since the initial instruction is underspecified, the Navigator engages in multi-turn dialog with a remote Guide to acquire additional guidance along the navigation. **Bottom:** At each step, the Navigator follows a modular decision process: it either proceeds autonomously (*Navigation*) or requests help (*Question Generation*). When a question is asked, the remote Guide localizes the Navigator and provides an answer describing the next path to the goal. This forms an alternating loop of dialog and action that continues until the goal region is reached.

the high complexity of dialog-based navigation. Yet collecting human annotated DialNav dataset is exceptionally expensive (*e.g.,* 7,500 USD for 2K episodes), as it requires two people to coordinate in real time while grounding their conversations. Although dataset augmentation techniques to mitigate the cost of human annotation have been explored in VLN (Fried et al., 2018; Fan et al., 2024; Wang et al., 2025b), they are restricted to single-turn settings and are not directly applicable to DialNav.

To address this data scarcity challenge, we propose an automatic DialNav episode generation pipeline. Our pipeline leverages existing fine-grained VLN datasets by concatenating their paths, generating scene captions to act as questions, and using an LLM to reformat these into natural dialog. Through this process, we yield the **RAINbow** dataset, which is more than two orders of magnitude larger than the original RAIN. Despite this substantial data scaling, our preliminary experiments reveal that data augmentation alone is not effective when paired with the existing training framework in Han et al. (2025). We identify that the training scheme from Han et al. (2025) fails to effectively leverage the augmented data. Therefore, we adopt Dual-Strategy Training, a training scheme better aligned with DialNav task that fully exploits our scaled multi-turn dialog-based episodes. We additionally improve the localization, an essential subtask to DialNav which remains underexplored.

Together, these advances improve performance of the holistic navigation-dialog loop by more than two-fold: success rate increases from 30.77 to 58.24 (+89%) on *Val Seen* and from 14.52 to 29.05 (+100%) on *Val Unseen*. We will publicly release our dataset, code, and models.

In summary, our contributions are threefold:

- We construct RAINbow, a large-scale dataset, expanding the training data by more than two orders of magnitude beyond RAIN.
- We adopt Dual-Strategy Training to better exploit large-scale dialog data and improve localization, an underexplored but critical subtask in DialNav.
- We deliver a new state of the art performance on DialNav, doubling success rate over the previous baseline.

2 Related Work

Vision and Language Navigation. Vision-and-Language Navigation (VLN) is an embodied AI task in which an agent navigates through visual environments by following natural language instructions. Instructions are either coarse-grained (Zhu et al., 2021a; Qi et al., 2020), often too ambiguous for reliable navigation, or fine-grained (Anderson et al., 2018; Ku et al., 2020; Chen et al., 2019), often overly detailed and unnatural for human interaction. Some prior works have explored incorporating dialog or interaction in navigation (Thomason et al., 2020; Roman et al., 2020; de Vries et al., 2018; Banerjee et al., 2021; Nguyen and Daumé III, 2019; Zhu et al., 2021b; Zhao et al., 2026; Taioli et al., 2025). The recently proposed

DialNav task (Han et al., 2025) places particular emphasis on the dialog through its remote guide setup and dialog-navigation loop evaluation. Although DialNav provides a valuable benchmark, its performance is limited primarily due to its small dataset size. In this work, we improve DialNav performance by large-scale data augmentation, complemented by enhancements in training.

VLN dataset augmentation. Data scarcity is a well-known challenge in VLN. To address this, various data augmentation approaches have been explored. Some works focused on training a text generation model to generate new instructions for unannotated paths (Fried et al., 2018; Zhang and Kordjamshidi, 2023; Wang et al., 2024a). Other works scaled environments, either by editing existing scenes (Tan et al., 2019; Liu et al., 2021; Li et al., 2022) or leveraging large sets of additional environments (Wang et al., 2023c; Chen et al., 2022a; Koh et al., 2022; Guhur et al., 2021; Lin et al., 2023; Wang et al., 2026). Recently, more advanced methods have emerged, adopting LLM generators or applying additional filtering or refinement (Zeng et al., 2023; Fan et al., 2024; Kong et al., 2024; Wang et al., 2025b, 2024b). However, these methods are exclusively focused on generating *single-turn instructions*. Our work proposes a simple yet powerful pipeline that converts existing single-turn VLN datasets into multi-turn DialNav episodes. Through this pipeline, we generate RAINbow dataset, which expands the original RAIN dataset (Han et al., 2025) by more than two orders of magnitude.

Embodied Dialog. Embodied dialog presents challenges in both dataset and training. To mitigate the challenge of data scarcity, some works explore data augmentation leveraging large language models (Padmakumar et al., 2023) or template-based procedures (Gao et al., 2022). For training, many existing works primarily train and evaluate models on static dialog histories (Thomason et al., 2020; Padmakumar et al., 2022; Hahn et al., 2020). Some works propose training schemes for holistic execution-dialog loop setups, using human correction (de Vries et al., 2018), reinforcement learning (Roman et al., 2020), or training with all possible dialog cases (Gao et al., 2022). The prior work on DialNav (Han et al., 2025) trains agents in a static setup, which creates a train-test mismatch. In this work, motivated by recent findings that large-scale synthetic training data can substantially improve model capabilities (Wang et al., 2024b; Min et al., 2025), we generate a high-quality, well-grounded dataset at low cost, and adopt a training scheme to fully utilize this large-scale data.

3 Preliminaries

Notation. In DialNav (Han et al., 2025), the *Navigator* agent traverses toward a goal with the assistance of a remote *Guide*, as illustrated in Figure 1 (top). Formally, the environment is represented as a connectivity graph $G = (V, E)$, where $v \in V$ is a navigable node, and $E \subseteq V \times V$ is the set of navigable edges between nodes. A DialNav episode $\mathcal{E}$ is defined as $\mathcal{E} = (G, b, R, I, T_J, D_J)$, where $b \in V$ is the initial node, $R \subseteq V$ is the goal region spanning multiple adjacent nodes, and I is an initial instruction. The Navigator's complete trajectory $\mathcal{T}$ is a sequence of traversed nodes over K steps $\mathcal{T} = (v_0, v_1, \ldots, v_K)$, where $v_0 = b$ and $v_K \in R$. During navigation, J-turn dialog occurs, denoted as $D_J = ((q_1, a_1), \ldots, (q_J, a_J))$, with each (q_j, a_j) exchanged at a dialog point $v_{d_j} \in \mathcal{T}$. These J dialog points segment the full trajectory $\mathcal{T}$ into $J + 1$ sub-trajectories: $(T^{(0)}, T^{(1)}, \ldots, T^{(J)})$, where $T^{(j)} = (v_{d_j}, \ldots, v_{d_{j+1}-1})$.

Navigator and Guide Workflow. The DialNav task unfolds as a holistic workflow integrating all subtasks (Figure 1 bottom). The episode begins with the Navigator at b with the initial instruction I, and the remote Guide with knowledge of the full environment G, the goal R, and I. The process then unfolds in rounds. At each step t, the Navigator first decides whether to ask for help. If the Navigator chooses to *proceed without help*, it autonomously selects its next action and advances to node v_{t+1}. If the Navigator chooses to *ask for help*, it generates a natural language question q_j, and its current node is denoted as a dialog point v_{d_j}. Upon receiving the question, the remote Guide first localizes the Navigator by estimating its dialog point as $\hat{v}_{d_j}$. Then, it generates a natural language answer a_j, typically describing the path from the inferred position $\hat{v}_{d_j}$ to R. The Navigator continues navigation using this exchange (q_j, a_j) as additional guidance. This process repeats until the Navigator determines it has reached the goal R.

4 Method

In this section, we first present our automatic episode generation pipeline (§4.1). This pipeline leverages existing VLN datasets to synthesize a large-scale dataset, **RAINbow**, expanding the available training data by over two orders of magnitude.

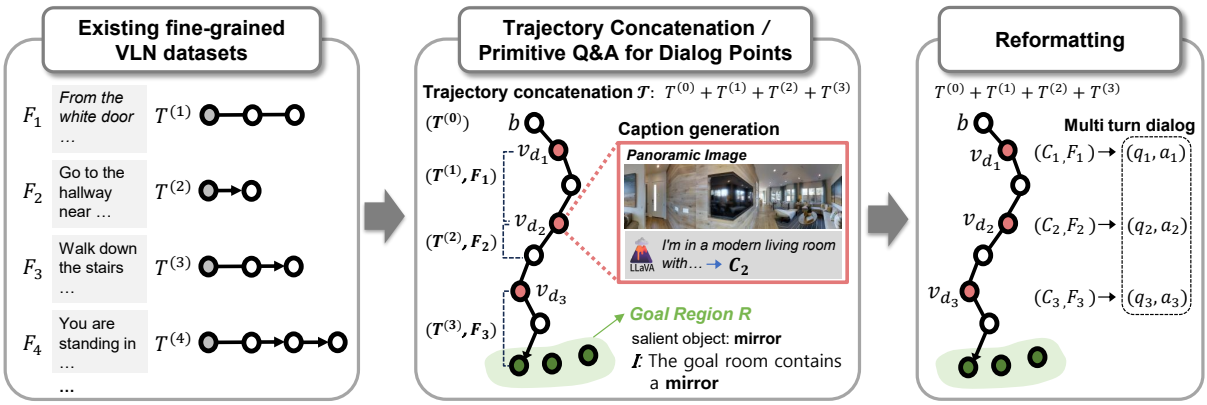


Figure 2: **Overview of the dataset generation pipeline. (Left)** We start from existing single-turn fine-grained VLN datasets, where each path is paired with its instruction F_j. **(Middle)** Multiple sub-trajectories are concatenated into an extended trajectory. The starting node of each sub-trajectory becomes a dialog point v_{d_j}, and at each dialog point, a panoramic caption C_j is generated using a vision–language model. The original fine-grained instructions F_j are repurposed as dialog answers, while the final node defines the goal region R. **(Right)** Caption–instruction pairs (C_j, F_j) are then reformatted into natural multi-turn dialogs using a large language model, producing large-scale dialog-style data for DialNav training.

We then introduce complementary enhancements to the navigation training (§4.2.1) and localization model (§4.2.2), enabling the model to better exploit this large-scale augmented dataset.

4.1 RAINbow Dataset Generation

Collecting dialog-based navigation datasets is prohibitively expensive, as it requires two people to coordinate in real time while grounding their conversations in the given task context. Consequently, the human-annotated RAIN dataset (Han et al., 2025) contains only 2K episodes, which is insufficient to capture the complexity of dialog-conditioned navigation and thus hinders generalization. To overcome this limitation, we automatically generate large-scale DialNav episodes by utilizing VLN datasets. We observe that, in DialNav (Han et al., 2025), dialog unfolds in turns along a long trajectory, where each question–answer exchange guides a subsequent portion of the path. To simulate this characteristic, we design our pipeline in three stages: (1) concatenating paths from existing VLN datasets; (2) generating scene captions for each dialog point (to serve as question content); and (3) rewriting the caption-instruction pairs into natural dialog using an LLM. Figure 2 illustrates the RAINbow data generation pipeline. We describe each stage in detail below.

Trajectory Concatenation. From three widely used VLN datasets, R2R (Anderson et al., 2018), RxR (Ku et al., 2020) and CVDN[1] (Thomason et al., 2020), we concatenate 2–4 paths to form a new full trajectory $\mathcal{T}$. Each path, which becomes a sub-trajectory $T^{(j)}$, is paired with its original fine-grained instruction F_j. We then form the initial exploration path $T^{(0)}$ by prepending up to five navigation steps to the start node of $T^{(1)}$. To better reflect the noise encountered in real dialog-based navigation, we additionally inject *mislocalization*, *misnavigation*, and *exploration* scenarios into 10% of the constructed trajectories and their corresponding dialogs. From the resulting trajectory $\mathcal{T} = (T^{(0)}, \ldots, T^{(J)})$, the initial node of $T^{(0)}$ becomes the episode's start node b, and the region containing the final node of $T^{(J)}$ becomes the goal region R. We then generate an initial instruction I that indirectly specifies the goal region R by describing an object located within it. This region and object information is based on the metadata in Chang et al. (2017). Further implementation details are provided in Appendix C.1.

Primitive Q&A for Dialog Points. Given the concatenated trajectory T, the initial node v_{d_j} of each sub-trajectory $T^{(j)}$(for $j \geq 1$) is treated as a *dialog point* where a dialog exchange occurs. In this step, we prepare the primitive content for the dialog exchange (q_j, a_j) at each dialog point v_{d_j}. Leveraging the question style of RAIN (Han et al., 2025), where questions typically describe the visual scene to aid localization, we first generate the question content, C_j, by captioning the observation at the dialog point via a large vision–language model (e.g., LLaVA-1.5-7B (Liu et al., 2023)). For the answer content, we repurpose the original fine-grained

[1]CVDN is a dialog dataset with answers describing the 5 next nodes along the shortest path to the goal, which we utilize as fine-grained instructions.

Figure 3: **Qualitative example of RAINbow.** This figure shows a 2-turn dialog episode in RAINbow. The left column shows the first dialog exchange, and the right column shows the second. The generated dialogs exhibit a natural conversational flow across turns and are well-grounded in both the Navigator's visual observations (*e.g.,* "fireplace", "pool table") and the route the Guide describes (*e.g.,* "leave this room", "past the pool table").

instruction F_j associated with the sub-trajectory $T^{(j)}$. The resulting (C_j, F_j) thus form primitive question–answer pairs aligned with the dialog point of the concatenated trajectory. As these raw pairs are not yet in natural dialog form and the flow across turns is often awkward, we refine them in the following step. Details are provided in the Appendix C.

Reformatting into Multi-turn Dialog. Finally, we reformat the caption–instruction pairs (C_j, F_j) obtained from the previous steps into coherent multi-turn dialogs. While the caption C_j describes the local scene and the instruction F_j gives a specific direction for the next sub-trajectory $T^{(j)}$, these sentences are originally written as independent sentences and therefore do not form a natural conversation. To bridge this gap, we employ an LLM (*e.g.,* GPT-4o-mini) to rewrite the raw (C_j, F_j) pairs into fluent question–answer exchanges (q_j, a_j). To reduce hallucination and ensure faithfulness to the intended goal, we first refine each instruction before converting it into dialog, and further validate this process through human evaluation and closed-source VLMs. Details of the prompting and validation are provided in the Appendix C.

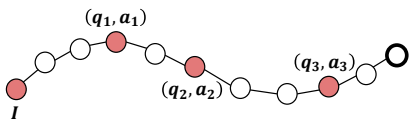
The RAINbow Dataset. Through this process, we synthesize the **RAINbow** (**RAIN** built **on** wide set) dataset. The resulting RAINbow is a 238K episode dataset that is over two orders of magnitude larger than the RAIN dataset. Figure 3 shows an example of the final generated dialog data. The key characteristics in RAINbow remain comparable to the RAIN dataset. In Rainbow, the average numbers of dialog turns, trajectory nodes, question words, and answer words are 2.71, 19.54, 30.43, and 43.17, respectively (compared to 1.88, 26.03, 27.48, and 42.53 in RAIN). The reformatting step using GPT-4o-mini incurred a total cost of about USD 400, equating to USD 0.0016 per episode. This makes our generation pipeline about 2,000 times more cost-effective than manual annotation in RAIN (USD 3.75 per episode). Statistics and more examples of RAINbow are provided in Appendix A.

4.2 Training the Navigator and Guide

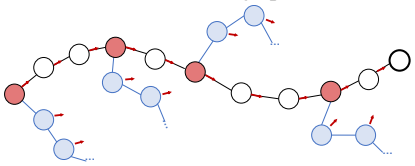
To train the Navigator and the Guide, we follow the modular design introduced in Han et al. (2025). As described in the workflow in Figure 1 (bottom), each agent is decomposed into task-specific modules—*whether-to-ask*, *question generation*, and *navigation* for the Navigator, and *localization* and *answer generation* for the Guide. For the whether-to-ask module, we adopt a confidence thresholding strategy, prompting the agent to initiate a question whenever the model confidence falls below a predefined threshold. For the question and answer generation modules, we follow the training scheme of Han et al. (2025).

4.2.1 Dual-Strategy Training for Navigation

The segment-based training in prior work (Han et al., 2025) trains the navigator without previous navigation or future dialog updates, not taking the evolving dialog nature of DialNav into account. While training on the full episode with dialog updates along the path seems to solve this, it presents a fundamental dilemma: the agent must strictly follow the annotated path to receive valid dialog updates (q_j, a_j) at the correct nodes v_{d_j}. However, this strict adherence prevents the agent from learn-



(a) DialNav training episode $\mathcal{E}$



(b) Dual-Strategy Training

Figure 4: **Comparison of navigation training strategies.** (a) **DialNav training episode** $\mathcal{E}$: An episode unfolds from an initial node (marked I), with dialog dynamically updated at each dialog points (red nodes). (b) **Dual-Strategy Training**: The *data-guided rollout* follows the annotated path in dataset (white nodes), updating dialog at each dialog point (red nodes). This is supplemented with *on-policy rollouts* (blue nodes) forked at each dialog point. The model is supervised to predict the next node along the dataset path for data-guided rollouts and the shortest path to the goal for on-policy rollouts, respectively (red arrows).

ing to recover from its own errors.

To this end, we adopt a **Dual-Strategy Training** for navigation that combines two types of rollouts: data-guided and on-policy (Figure 4b). The *data-guided rollout* (Figure 4b, white nodes) constrains the Navigator to follow the dataset trajectory, allowing it to arrive at each dialog node v_{d_j} and receive the ground-truth dialog exchange (q_j, a_j) as a dynamic instruction update. This is complemented by an *on-policy rollout* (Figure 4b, blue nodes), forked at each dialog point v_{d_j} (Figure 4b, red nodes), where the agent follows its own policy without receiving further dialog updates, supervised to predict the next node on the shortest path to the goal region R (Figure 4b, red arrows). By jointly training on both data-guided and on-policy rollouts, the navigation module learns to leverage the full episode while remaining robust to its own prediction errors.

The loss for each dialog point, v_{d_j} is defined as follows:

$$\mathcal{L} = \underbrace{\sum_{i=d_j}^{d_{j+1}-1} \log p(v_{i+1}|v_{1:i}, D_j)}_{\text{data-guided rollout}} + \underbrace{\sum_{t=d_j}^{K-1} \log p(\hat{v}_{t+1}|v_{1:d_j}, \bar{v}_{d_j+1:t}, D_j)}_{\text{on-policy rollout}}, \quad (1)$$

in which $\hat{v}$ is the next step on the shortest path and $\bar{v}$ denotes the next step sampled from the model's own policy of the navigation model.

4.2.2 Transferring VLN Knowledge for Localization

Localization is a subtask arising from DialNav's remote Guide setup. The Guide, who knows the entire map G, must predict the dialog point v_{d_j} based on the Navigator's question q_j to provide accurate guidance. Localization tasks share similarities with VLN in their environment and task, as both operate in the Matterport3D (Chang et al., 2017) and are text-grounded node selection tasks. This suggests that localization can benefit from models and pretraining strategies originally developed for VLN (Hahn and Rehg, 2022). Motivated by this, we adopt the graph-based Transformer architecture of Chen et al. (2022b) for localization. As shown in our experiments, this significantly improves localization performance. Details of the model are provided in Appendix E.

5 Experiments

5.1 Experimental Settings

Implementation of Navigator and Guide. Following prior work (Han et al., 2025), we use DUET (Chen et al., 2022b) for *navigation* and LANA (Wang et al., 2023a) for the *question/answer generation* modules. For the *localization* module, we compare the GCN from Han et al. (2025) with Graph-based Transformer Localization (§4.2.2). All models are initialized from publicly available pre-trained weights. We then train the modules using a 1:9 mixture of RAIN and RAINbow data. For the *whether-to-ask* module, we fix the confidence threshold of 0.9. Details of each model architecture are provided in Appendix E.

Evaluation. We evaluate our method on the RAIN validation dataset, which is split into *Val Seen* and *Val Unseen* (Anderson et al., 2018). The *Val Seen* split consists of environments seen during training, albeit with novel paths, whereas *Val Unseen* contains entirely novel environments. We report performance under the holistic dialog–navigation loop setting, where the Navigator starts with an initial instruction and must proactively ask questions to receive guidance and reach the goal. We measure performance using standard VLN metrics: Success Rate (SR), Oracle Success Rate (OSR), Success weighted by Path Length (SPL), and Navigation Error (NE, in meters), which capture navigation

#	Setup	Val Seen						Val Unseen					
		SR↑	OSR↑	SPL↑	NE↓	NSC	DTC	SR↑	OSR↑	SPL↑	NE↓	NSC	DTC
(1)	Baseline (Han et al., 2025)	30.77	38.46	26.93	9.76	18.65	3.44	14.52	19.50	11.33	16.38	15.52	5.61
(2)	+RAINbow	34.07	43.96	32.21	9.61	19.64	5.34	18.26	21.58	13.89	16.27	23.00	13.76
(3)	+DST	38.46	45.05	32.78	8.92	22.58	6.38	15.77	24.90	11.79	16.92	21.29	10.16
(4)	+RAINbow+DST	50.55	60.44	45.60	6.22	22.13	5.63	25.73	**32.78**	19.13	14.49	23.85	13.30
(5)	**+RAINbow+DST+GTL (Ours)**	**58.24**	**68.13**	**51.65**	**5.38**	21.30	4.66	**29.05**	31.54	**19.61**	**13.91**	21.46	11.08

Table 1: **Performance gains from progressively adding components. +RAINbow**: training with RAINbow dataset (§4.1). **+DST**: training navigation module with Dual-Strategy Training (§4.2.1). **+GTL**: adopting Graph-based Transformer Localization model (§4.2.2). RAINbow alone (Row 2) offers performance gains, but its potential is unlocked when combined with our proposed DST (Row 4), which significantly amplifies its effect. GTL (Row 5) provides further improvement, leading to the best overall performance.

#	Dataset	Val Seen						Val Unseen					
		SR↑	OSR↑	SPL↑	NE↓	NSC	DTC	SR↑	OSR↑	SPL↑	NE↓	NSC	DTC
(1)	RAIN	37.36	47.25	33.74	8.54	22.16	6.01	23.24	29.46	16.49	14.68	20.80	8.85
(2)	+ VLN instructions	40.66	51.65	32.73	8.94	22.93	9.56	25.73	**34.44**	17.92	14.37	24.68	12.67
(3)	+ Traj. Concat	39.56	48.35	35.36	7.46	20.54	4.00	19.09	27.80	13.79	17.04	25.80	13.06
(4)	+ Primitive Q&A	56.04	65.93	51.41	5.51	20.22	5.80	22.82	**34.44**	15.47	14.32	23.15	12.19
(5)	**+ Reformat (RAINbow)**	**58.24**	**68.13**	**51.65**	**5.38**	21.30	4.66	**29.05**	31.54	**19.61**	**13.91**	21.46	11.08

Table 2: **Dataset augmentation ablations.** RAIN: training only with the RAIN dataset. **+ VLN instructions**: adding VLN instructions. **+ Traj. Concat**: concatenating trajectories from VLN instructions. **+ Primitive Q&A**: using primitive QA content. **+ Reformat**: reformatting into natural multi-turn QA dialogs.

quality and efficiency. We also report Navigation Step Count (NSC) and Dialog Turn Count (DTC) to measure exploration behavior and dialog usage.

5.2 Results

Impact of the proposed components. Table 1 presents the impact of our proposed components: adding the RAINbow dataset, Dual-Strategy Training (DST), and Graph-based Transformer Localization (GTL). RAINbow alone (Row 2) provides a modest improvement over the baseline (Row 1) in SR under both setups (Val Seen +3.30, Val Unseen +3.74). However, its true potential is realized when paired with DST (Row 4), which yields substantially larger gains (Val Seen +12.09, Val Unseen +9.96, compared to Row 3). This result confirms our hypothesis: while the baseline segment-based training (Han et al., 2025) cannot fully leverage RAINbow, DST effectively utilizes the large-scale, multi-turn episodes and unlocks their full potential. Finally, adding GTL (Row 5) produces the best performance: SR reaches 58.24 (+89%) on Val Seen and 29.05 (+100%) on Val Unseen, roughly doubling the baseline success rate on both splits.

Ablations of the Data Augmentation Pipeline. Table 2 analyzes the effect of each stage in our augmentation pipeline. For fair comparison, we applied Dual-Strategy Training and Graph-based Transformer Localization across all setups. Naively adding publicly available fine-grained VLN instructions (Anderson et al., 2018; Ku et al., 2020) (Row 2), providing a baseline for data augmentation, improved navigation SR +3.30 (+9%) in Val Seen and +2.49 (+11%) in Val Unseen. This gain from simply exposing the Navigator to more environments highlights the importance of data diversity for improving performance. With trajectory concatenation (Row 3), the performance drops compared to Row 2. We attribute this to data reduction; the constraints for connecting paths (§4.1) necessarily discard some source trajectories, providing the navigator with less environmental coverage than the full VLN instruction set. For Rows 2 and 3, only the navigation module can use the augmented data, because these settings lack the dialog content required to train the localization, question generation, and answer generation modules. Introducing Primitive Q&A (Row 4) addresses this issue by providing supervision across all modules, resulting in a substantial improvement in SR on Val Seen: +18.68 (+50.00%) compared to Row 1. However, this setting still struggles in Val Unseen. We conjecture this is due to a style mismatch, as modules are trained on a mixture of human-annotated RAIN dialogs and the unformatted, raw content from the pipeline, which hinders generalization. Finally, Reformat (Row 5) converts this raw content into natural, contextually coherent dialogs, yielding the best results (Seen SR: 58.24, Unseen SR: 29.05). Crucially, reformatting significantly boosted SR in Val Unseen (compared to Row 4, +6.23). Over-

#	Training Method	SR↑	SPL↑	NSC	DTC
Val Seen					
(1)	Han et al. (2025)	32.97	30.76	19.85	5.76
(2)	**DST**	**58.24**	**51.65**	21.30	4.66
(3)	– On-policy Rollout	**58.24**	50.05	20.80	6.21
Val Unseen					
(4)	Han et al. (2025)	22.82	17.00	20.89	10.57
(5)	**DST**	**29.05**	**19.61**	21.46	11.08
(6)	– On-policy Rollout	21.16	15.38	21.41	8.34

Table 3: **Navigation training methods ablation.** Comparison of baseline and Dual-Strategy Training (DST) with and without On-policy Rollout.

Detour ratio	1.0–1.3	≥ 1.3	Total
DST	11.20	**17.85**	**29.05**
– On-policy Rollout	10.79	10.37	21.16

Table 4: **Detour analysis on Val Unseen.** Comparison of Success Rate between DST and the variant without on-policy rollout across different detour ratio.

all, these results demonstrate that while raw VLN fine-grained data or naively concatenated trajectories provide limited benefits, our pipeline's full process of structuring and naturalizing content into consistent dialog is effective for DialNav.

Ablations of Navigation Training. Table 3 ablates Dual-Strategy Training (§4.2.1) on both the Val Seen and Val Unseen splits. For fair comparison, we applied the RAINbow data and Graph-based Transformer Localization across all setups. Comparing Dual-Strategy Training (DST) (Row 2 and 5) to the Baseline (Row 1 and 4), we see a dramatic improvement in SR across both splits (Val Seen +76.6%, Val Unseen +27.3%). This confirms that Dual-Strategy Training (DST) that leverages the full, dynamic dialog loop is critical. Row 3 and 6 ablates our training by removing the on-policy rollout. While this maintains high SR in Val Seen, there is a severe drop in Val Unseen (-7.89). Table 4 further analyzes SR by detour ratio (path length / shortest path) on Val Unseen. When the detour is small (1.0–1.3), DST and the variant without on-policy rollout achieve comparable SR (11.20 vs. 10.79). However, when the detour is larger ($\geq$ 1.3), DST yields substantially higher SR (17.85) than the variant without on-policy rollout (10.37). By explicitly incorporating exploration through on-policy rollout during training, our model learns to recover from trajectory errors and make more robust predictions.

Ablations of Localization Model. Table 5 ablates the Graph-based Transformer Localization model. To isolate the localization performance, we evaluate it outside the holistic dialog-navigation loop. Instead, we use the human-annotated questions from the validation set (the human-human dialog for validation episodes in RAIN); the model is tasked to predict the node from which that question was asked. We report Localization Error (LE) in meters and 3m accuracy (A@3m), the percentage of predictions within 3m of the ground truth. Compared to the GCN baseline from prior work (Han et al., 2025) (Row 1), changing the architecture to Graph-based Transformer Localization alone (Row 2) degrades performance, highlighting the difficulty of training this complex model from scratch on scarce data. However, initializing it with VLN pretrained weights (Row 3) yields a dramatic improvement, substantially surpassing the baseline (Row 1) in both environments. This confirms the efficacy of transferring VLN knowledge for localization. These results demonstrate that GTL, enhanced by knowledge transfer and the RAINbow dataset, effectively improves localization.

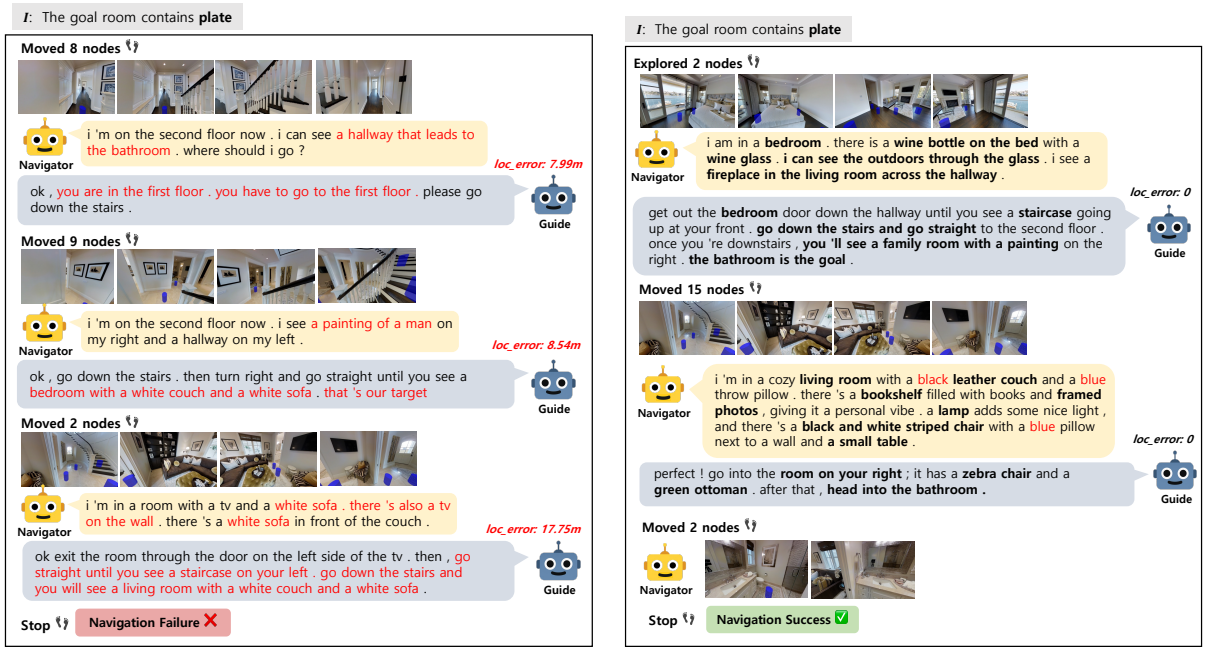
#	AC	VT	Val Seen		Val Unseen	
			LE↓	A@3m↑	LE↓	A@3m↑
(1)			10.72	40.70	14.47	21.74
(2)	✓		17.42	16.37	16.25	15.31
(3)	✓	✓	**8.16**	**49.12**	**11.45**	**34.97**

Table 5: **Localization model ablation. AC**: Architecture change to Graph-based Transformer Localization. **VT**: VLN pretrained knowledge transfer. All together, the new localization model reduces localization error (LE) by 2.56m in Val Seen and 3.02m in Val Unseen compared to the baseline.

Agent Role	Method	Val Seen		Val Unseen	
		SR ↑	HS ↑	SR ↑	HS ↑
Navigator	Baseline	39.1	2.89	13.6	2.16
	Ours	**43.5**	**3.18**	**20.0**	**2.24**
Guide	Baseline	47.8	3.18	14.3	2.10
	Ours	**73.1**	**3.75**	**44.0**	**2.52**

Table 6: **Human-agent cooperation results on Dial-Nav. SR**: Navigation Success Rate (%). **HS**: Human Score, representing the perceived helpfulness of the agent counterpart, rated by human participants on a scale from 1 to 5. For each setting, either the human or the agent takes the Guide role, while the other takes the Navigator role.

Human-Agent Cooperation. To further validate the effectiveness of our approach, we evaluate it in a human–agent cooperation setup (Table 6). We build a simulator in which a human participant collaborates with an agent: one takes the role of



(a) Baseline (Han et al., 2025)

(b) Ours

Figure 5: **Qualitative comparison on the same task instance between the baseline (Han et al., 2025) (left) and Ours (right).** The baseline agent produces broken language with wrong details (marked in red), likely due to dataset scarcity, leading to high localization errors and navigation failure. In contrast, our agent provides richer, well-grounded descriptions (marked in **bold**), yielding accurate localization and reliable instructions, ultimately leading to successful navigation.

the Guide, while the other acts as the Navigator. Our method consistently outperforms the baseline across all roles and environments, demonstrating that it substantially improves human–agent collaboration. Details are provided in Appendix Sec J.

Qualitative Examples. Figure 5 and Appendix Sec. H provide qualitative comparisons between the baseline Navigator–Guide agent pair (Han et al., 2025) and ours.

6 Conclusion

In this work, we addressed the key limitations of DialNav, a holistic dialog-based navigation task whose performance has been constrained by critical data scarcity. We first proposed an automatic data generation pipeline to create RAINbow, a large-scale (238K) dataset that expands the available training data by utilizing existing VLN instructions, addressing the core data scarcity issue. Second, we adopted the Dual-Strategy Training scheme for the navigator, aligning the training process with the dynamic, multi-turn nature of the task by combining data-guided rollout and on-policy rollout. Finally, we adapted a graph-based transformer for the localization, enabling us to leverage powerful VLN pretraining for the localization task. Our experiments demonstrated that these components are highly synergistic. Integrating these solutions, our final model establishes a new state-of-the-art, more than doubling the baseline success rate on both *Val Seen* (30.77 to 58.24, +89%) and *Val Unseen* (14.52 to 29.05, +100%) splits. By providing a large-scale, high-quality dataset and adopting training and model to leverage it, this work lays a stronger foundation for future research into more complex, interactive embodied agents.

Limitations

While our work substantially improves DialNav through large-scale automatic data generation and training enhancements, several limitations remain. Although RAINbow significantly increases data scale, it is built upon existing VLN trajectories, which may limit environmental and behavioral diversity. In addition, the generated dialogs, while generally natural and goal-consistent, may not fully capture the full richness of human embodied communication. Our evaluation is primarily conducted in indoor navigation settings, and further validation in broader embodied domains would be beneficial.

Acknowledgements

This research was supported by MSIT (IITP: IITP-2026-RS-2024-00436857, RS-2024-00398115, IITP-2026-RS-2020-II201819, IITP-2026-RS-2025-02304828, RS-2026-25507282, RS-2026-25585074; NRF: RS-2025-23523979), MCST (KOCCA: RS-2026-25506607, RS-2024-00345025), MND (RS-2026-25551943) and MSS (RS-2026-25549378). This research was also supported by the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (MSIT).

References

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

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

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12538–12547.

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

Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. 2022b. 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.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue. *Preprint*, arXiv:1807.03367.

Sheng Fan, Rui Liu, Wenguan Wang, and Yi Yang. 2024. Navigation instruction generation with bev perception and large language models. In *European Conference on Computer Vision*, pages 368–387. Springer.

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

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

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

Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. 2021. Airbert: In-domain pretraining for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1634–1643.

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

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

Leekyeung Han, Hyunji Min, Gyeom Hwangbo, Jonghyun Choi, and Paul Hongsuck Seo. 2025. Dialnav: Multi-turn dialog navigation with a remote guide. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8514–8523.

Julia Kiseleva, Ziming Li, Mohammad Aliannejadi, Shrestha Mohanty, Maartje ter Hoeve, Mikhail Burtsev, Alexey Skrynnik, Artem Zholus, Aleksandr Panov, Kavya Srinet, and 1 others. 2022. Interactive grounded language understanding in a collaborative environment: Iglu 2021. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 146–161. PMLR.

Jing Yu Koh, Harsh Agrawal, Dhruv Batra, Richard Tucker, Austin Waters, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. 2022. Simple and effective synthesis of indoor 3d scenes. In *AAAI Conference on Artificial Intelligence*.

Xianghao Kong, Jinyu Chen, Wenguan Wang, Hang Su, Xiaolin Hu, Yi Yang, and Si Liu. 2024. Controllable navigation instruction generation with chain of thought prompting. In *European Conference on Computer Vision*.

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

Jialu Li, Hao Tan, and Mohit Bansal. 2022. Envedit: Environment editing for vision-and-language navigation. In *Computer Vision and Pattern Recognition*.

Kunyang Lin, Peihao Chen, Diwei Huang, Thomas H Li, Mingkui Tan, and Chuang Gan. 2023. Learning vision-and-language navigation from youtube videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8317–8326.

Chong Liu, Fengda Zhu, Xiaojun Chang, Xiaodan Liang, and Yi-Dong Shen. 2021. Vision-language navigation with random environmental mixup. In *IEEE International Conference on Computer Vision*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Hyunji Min, Sangwon Jung, Junyoung Sung, Dosung Lee, Leekyeung Han, and Paul Hongsuck Seo. 2025. Goat: A training framework for goal-oriented agent with tools. *arXiv preprint arXiv:2510.12218*.

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

Aishwarya Padmakumar, Mert Inan, Spandana Gella, Patrick Lange, and Dilek Hakkani-Tur. 2023. Multimodal embodied plan prediction augmented with synthetic embodied dialogue. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6114–6131, Singapore. Association for Computational Linguistics.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. Teach: Task-driven embodied agents that chat. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2017–2025.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020. 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.

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

Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. 2019. Executing instructions in situated collaborative interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2119–2130.

Francesco Taioli, Edoardo Zorzi, Gianni Franchi, Alberto Castellini, Alessandro Farinelli, Marco Cristani, and Yiming Wang. 2025. Collaborative instance object navigation: Leveraging uncertainty-awareness to minimize human-agent dialogues. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18781–18792.

Hao Tan, Licheng Yu, and Mohit Bansal. 2019. Learning to navigate unseen environments: Back translation with environmental dropout. In *North American Chapter of the Association for Computational Linguistics*.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2020. Vision-and-dialog navigation. In *Conference on Robot Learning*, pages 394–406. PMLR.

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

Liuyi Wang, Zongtao He, Ronghao Dang, Mengjiao Shen, Chengju Liu, and Qijun Chen. 2024a. Vision-and-language navigation via causal learning. In *Computer Vision and Pattern Recognition*.

Xiaohan Wang, Wenguan Wang, Jiayi Shao, and Yi Yang. 2023a. 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.

Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, and 1 others. 2023b. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20270–20281.

Zihan Wang, Seungjun Lee, Yinghao Xu, and Gim Hee Lee. 2026. Image2sim: Scaling embodied navigation via generative neural simulator. *arXiv preprint arXiv:2607.05765*.

Zihan Wang, Yaohui Zhu, Gim Hee Lee, and Yachun Fan. 2025b. Navrag: Generating user demand instructions for embodied navigation through retrieval-augmented llm. *arXiv preprint arXiv:2502.11142*.

Zun Wang, Jialu Li, Yicong Hong, Songze Li, Kunchang Li, Shoubin Yu, Yi Wang, Yu Qiao, Yali Wang, Mohit Bansal, and 1 others. 2024b. Bootstrapping language-guided navigation learning with self-refining data flywheel. *arXiv preprint arXiv:2412.08467*.

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

Haitian Zeng, Xiaohan Wang, Wenguan Wang, and Yi Yang. 2023. Kefa: A knowledge enhanced and fine-grained aligned speaker for navigation instruction generation. In *arXiv.org*.

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

Yue Zhang and Parisa Kordjamshidi. 2023. Vln-trans: Translator for the vision and language navigation agent. In *Annual Meeting of the Association for Computational Linguistics*.

Xunyi Zhao, Sihao Lin, Gengze Zhou, Zerui Li, Shijie Li, Wei Tao, Jiajun Liu, and Qi Wu. 2026. Ask when it pays: Cost-aware open-ended interaction for instance goal navigation. *arXiv preprint arXiv:2606.03175*.

Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, and Xiaodan Liang. 2021a. 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.

Yi Zhu, Yue Weng, Fengda Zhu, Xiaodan Liang, Qixiang Ye, Yutong Lu, and Jianbin Jiao. 2021b. 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.

A RAINbow Dataset

A.1 Qualitative Examples

We present qualitative examples from the RAINbow dataset in Figures 8 and 9. The dialogs in RAINbow are **well-grounded** in the environment and the interactions exhibit a **natural flow**. The multi-turn exchanges maintain contextual coherence and conversational fluency, closely mimicking the dynamics of human-to-human collaboration.

A.2 RAIN vs. RAINbow

We compare the original **RAIN** dataset (2,231 episodes) with our automatically generated **Rainbow** dataset (238,028 episodes). Table 7 reports mean, median, maximum, and minimum values across episodes for key statistics. Figure 6 compares dialog style between human annotated RAIN dataset and generated RAINbow dataset.

	RAIN			**Rainbow (Ours)**		
Statistic	Mean	Min	Max	Mean	Min	Max
# Episodes		2,231			238,028	
Dialog turns / ep.	1.88	0	8	2.71	1	4
Question length (words)	27.48	1	119	30.43	4	125
Answer length (words)	42.53	1	184	43.17	0	300
Trajectory length (nodes)	26.03	4	116	19.54	2	67

Table 7: **Comparison of dataset statistics between RAIN and Rainbow.** Rainbow is more than two orders of magnitude larger, with richer dialogs and more diverse language.

A.3 CVDN vs. RAINbow

We provide a qualitative comparison with CVDN. Their interaction dynamics differ due to task setups: CVDN (Guide knows Navigator's position) vs. DialNav (Guide must infer Navigator's location). In CVDN, the Navigator's questions often serve as a mere trigger for the next instruction (*e.g.,* "hello", "where now?"), and each QA turn tends to resemble a sequence of independent *instruction-following* instances. Conversely, DialNav requires the Navigator to share context for localization. This involves potential misalignments, such as mislocalization or misnavigation, which RAINbow captures in its pipeline (details in Supp. Sec. C.1.7) exhibiting *decision-driven* actions, such as re-querying to resolve uncertainty.

RAINbow	**CVDN**
N: I'm in a bathroom with green tile walls and a painting over the ... *G:* Got it! From the washroom, turn a bit right towards the drawers and then ... *N: I think I might have gone off track a little.* Right now, I'm in a bright ... *G:* No problem, thanks for letting me know. Head out the open door next ...	*N:* hello *G:* hi, you need to leave the room and go to the room that has four two-seat sofas in it. *N:* where now? *G:* go up the stairs located beside the table in the room you are at now.

B Comparison of Embodied Dialog Dataset

Table 8 presents a comparison between RAINbow and existing embodied dialog datasets. Prior datasets typically rely on direct human annotation, which creates a trade-off between quality and cost, resulting in limited scale. In contrast, RAINbow is constructed by repurposing existing high-quality, human-annotated VLN datasets. By employing a simple yet effective pipeline to reformat these resources into dialog, we achieve a massive scale-up at a negligible cost per episode.

C Data Augmentation Details

In this section, we describe our scheme to generate realistic, visually well-grounded, and natural-flowing dialog in detail. Our pipeline is designed around three stages: (1) trajectory concatenation, (2) detailed caption generation for each node, and (3) reformatting into a natural dialog flow. This process ensures the resulting dialog is contextually grounded while preserving the fine-grained details from the well-formed captions and original VLN descriptions.

C.1 Trajectory Concatenation Details

For trajectory concatenation, we applied the following constraints to ensure natural continuity:

1. We concatenated 2-4 trajectories from R2R (Anderson et al., 2018), RxR (Ku et al., 2020), CVDN (Thomason et al., 2020) answer trajectories into a single episode. For CVDN dataset, we use its answer and pair its path with the next 5 shortest path from the question node based on its data collection description.

2. The endpoint of one trajectory and the start of the next must be within 1 meter in the navigation graph.

3. To prevent overly circuitous paths, the detour ratio, the concatenated path length divided by the shortest path length between the start and end nodes, was constrained to be less than 1.3.

Dataset	Task	Size	Total Cost	Cost/Ep	Source
CerealBar (Suhr et al., 2019)	Game	1K	$5.8K	$5.80	Human
IGLU (Kiseleva et al., 2022)	Game	8.9K	-	-	Human
HoloAssist (Wang et al., 2023b)	Manipulation	2K	-	-	Human
TEACh (Padmakumar et al., 2022)	Manipulation	4K	$105K	$26.25	Human
DialFRED (Gao et al., 2022)	Manipulation	53K	$10K	$0.19	Hybrid
RobotSlang (Banerjee et al., 2021)	Navigation	0.2K	-	-	Human
TalkTheWalk (de Vries et al., 2018)	Navigation	10K	-	-	Human
AVDN (Fan et al., 2023)	Navigation	3K	-	-	Human
CVDN (Thomason et al., 2020)	Navigation	2K	$7K	$3.50	Human
RAIN (Han et al., 2025)	Navigation	2K	$7.5K	$3.75	Human
RAINbow (Ours)	**Navigation**	**238K**	**$0.2K**	**< $0.01**	**Auto**

Table 8: **Comparison with existing embodied dialog datasets.** Most existing datasets relying on human annotation are costly, limited in scale, or restricted to synthetic environments. **Source**: 'Human' denotes human-annotated datasets, while 'Auto' denotes automatically generated ones.

4. Episodes in which the goal region contained no selectable object were discarded.

5. The ambiguous instruction I was derived from the Matterport3D (Chang et al., 2017) metadata by randomly selecting one visible object in the goal region. To avoid overly generic references, we excluded a predefined set of objects (e.g., `wall`, `floor`, `ceiling`, etc.).

6. Since goal regions in DialNav correspond to rooms rather than single nodes, we excluded cases where the agent had already reached the goal room before subsequent dialog turns were appended, avoiding unnatural "post-goal" interactions.

7. To further increase diversity, we additionally introduced variations in 10% of the constructed episodes, simulating potential deviations in real dialog navigation. We consider three types of variations: *mislocalization*, *misnavigation*, and *exploration*. In the case of mislocalization, the Guide intentionally provides an incorrect path description that does not match the Navigator's true position; the Navigator then proceeds by moving 1–2 nodes randomly from the original location rather than following the erroneous instruction. For misnavigation, the Navigator deviates from the instructed path and follows a randomly chosen alternative route. In exploration, the Navigator continues beyond the instructed trajectory by taking an additional 2–5 random steps after completing the suggested path. By incorporating these cases, we were able to include instructions that were previously filtered out by the strict trajectory-connection criteria, thereby maximizing the utilization of existing VLN data for dialog augmentation.

C.2 Caption Generation Details

To produce natural and visually grounded questions, we prompted LLaVA-1.5-7B (Liu et al., 2023) for visual description and Llama-3.1-8B for text synthesis with different visual/textual contexts. Our aim was to ensure that the resulting utterances contained sufficient local detail for Guide-side localization, while generating diverse content even for the same node. We experimented with three variants and randomly used one of these schemes for caption generation for each episode. The prompt for each variant is provided in Figure 10 and the resulting captions are provided in Figure 11.

(A) Simple panoramic caption: The panoramic observation at a node is directly provided to LLaVA with a short instruction to generate a navigation-oriented question. This produces concise questions.

(B) Region-grounded caption: Here, we first get the room type of each node from Matterport3D metadata. Then, we instructed LLaVA to provide a description of the panoramic scene, explicitly mentioning the room type. The room types are provided in Matterport3D metadata paired to each navigation node. This yields richer, more descriptive questions.

(C) Region and object-grounded panoramic caption: For a more detailed and visually grounded caption, we first generate captions for visible objects in the given node. Based on Matterport3D metadata, we first extract image patches of all visible objects at a given node. Then, we prompt LLaVA to generate a description of the given object image. To avoid hallucination or errors from wrong metadata annotations, we prompt it to gen-

erate 'none' if the given object is not visible in the image. We then jointly supply three inputs to LLaMA-3.1-8B: a general panoramic caption and a randomly sampled subset of the valid object-level captions. LLaMA is then prompted to generate a single detailed caption that integrates both global layout and salient object cues. To increase diversity, we vary the sampling ratio and the number of in-context examples. For in-context samples, we find similar samples in the RAIN dataset based on room types and visible objects.

C.3 Dialog Reformat Details

The primitive Q&A content, derived from raw captions and VLN instructions, is less natural than human conversation. To reformat this content into a natural multi-turn dialog, we employ GPT-4o-mini. To ensure formatting consistency and a natural flow, we perform this in two steps. The prompts for these steps are provided in Figure 12. The results of each steps are provided in Figure 13. Through this 2 step reformatting, the resulting dialog is natural, visually well grounded and better reflects realistic Navigator-Guide interactions.

1. **Instruction Refining.** First, we refine the raw VLN instructions. We prompt the LLM to make each instruction concise while preserving key spatial references, preventing overly verbose text that is unnatural in conversation. Moreover, since the original VLN instructions are for single-turn tasks, they often prematurely mention reaching the final goal. To prevent this, we also instruct the LLM to generate two versions of each refined instruction: a 'goal-oriented' variant (e.g., "You have reached your destination") and a 'neutral' variant (which does not mention the goal). We use the goal-oriented variant only for the final dialog turn and the neutral version for all intermediate turns. This explicit branching strategy prevents ambiguous or inconsistent goal mentions.

2. **Conversational Smoothing.** We then construct a draft dialog by sequencing the original scene captions and their corresponding refined, goal-conditioned instructions (from Step 1). This entire sequence is then paraphrased by GPT-4o-mini into fluent, conversational language, resulting in the final multi-turn dialog. This step retains all navigation-critical details while simplifying overly formal phrases, removing redundancies, and inserting subtle acknowledgments where appropriate to enhance conversational flow.

Split	# Objects	# Hallucination	Ratio
Questions	4.74	0.58	10.7%
Answers	4.77	0.13	2.0%

Table 9: **Average number of objects and hallucinations per turn.**

C.4 Dataset Evaluation.

To validate our final dataset we conducted a human evaluation. 30 samples each from RAIN and RAINbow were randomly shuffled and evaluated by 6 annotators. **(1) Goal Alignment:** When verifying if the dialog remains faithful to the initial goal, RAINbow achieved 90.0% accuracy (comparable to RAIN's 93.3%). This confirms that our *Instruction Refining* step (Supp. Sec. C.3.1) successfully preserves goal consistency. **(2) Naturalness:** In terms of linguistic flow, RAINbow achieved a score of 4.76/5.0, closely mirroring the human-written RAIN baseline 4.83/5.0.

C.5 Hallucination Analysis

To further assess the reliability of our pipeline, we conduct a hallucination evaluation on a sampled subset of RAINbow.

Evaluation protocol. We randomly sampled 100 episodes (247 dialog turns in total) and used a closed-source VLM (Claude Opus 4.7) to verify each object mentioned in the dialog against the actual observation. The Table 9 reports the average number of hallucinated objects per-turn.

Results. Our pipeline directly repurposes the human-annotated VLN instructions as answers, which results in low hallucination ratio (2.0%). The hallucination rate in questions is 10.7%. This low hallucination rate confirms the effectiveness of our pipeline.

D Results on CVDN

We evaluate our agents on the established Vision-and-Dialog Navigation (CVDN) dataset (Thomason et al., 2020). Table 10 presents a comparison against methods evaluated under the dialog in a loop setup. The CVDN evaluation uses a simplified remote Guide configuration: the Guide is provided with the Ground Truth localization and limits its guidance to 5 future navigation steps. Our

Method	Val Seen (m)	Val Unseen (m)
Shortest Path	32.8	29.3
$RMM_{N=3}$ (Roman et al., 2020)	14.0	5.6
SCoA (Zhu et al., 2021b)	19.5	11.2
Ours	**26.5**	**15.7**

Table 10: **Evaluation on the CVDN Dialog Setup.**The table compares our method against established baselines on the CVDN dataset (Thomason et al., 2020). The performance is **Goal Progress (m)**, where higher values indicate better performance.

approach achieves superior performance over all listed methods, demonstrating the strong generalization effect of our large-scale RAINbow dataset and training scheme to this external domain.

E Model Details

E.1 Agent Capabilities and Information Flow

The overall DialNav task relies on the specialized capabilities and distinct information access of the two agents. The integrated agents (Navigator and Guide) possess the full knowledge boundaries listed below. **Navigator Agent** (Navigation, Question, Whether-to-ask Decision): The Navigator has access to the accumulated dialog history, its navigational history (the subgraph of visited nodes and adjacent locations), and the initial instruction (I). **Guide Agent** (Localization, Answer): The remote Guide has knowledge of the full environment graph (G), the initial instruction (I), and the goal region (R). Since the Guide knows the full environment G, it can calculate the shortest path between any two nodes. However, in this work, for implementation and training convenience, each submodule utilized only the specific input features that its model architecture was designed to accept.

E.2 Navigation Model

For the core action prediction module (Navigator), we adopt the Dual-scale graph Transformer (**DUET**) (Chen et al., 2022b) architecture, following established practices in DialNav research (Han et al., 2025). DUET is a strong VLN backbone designed for joint long-term action planning and fine-grained cross-modal understanding.

Architecture and Reasoning. DUET explicitly builds a topological map and dynamically combines two scales of encoding via graph transformers: a fine-scale encoding of local observations and a coarse-scale encoding of the global map. This dual-scale approach allows the Navigator to reason efficiently over a large action space.

Pretraining. We leverage publicly available weights from large-scale VLN pretraining frameworks, such as ScaleVLN (Wang et al., 2023c). This initialization is vital because the pretraining process exposes the model to millions of diverse instruction-trajectory pairs across various environments, imparting robust environmental knowledge that is crucial for significantly reducing the generalization gap between navigating in seen and unseen environments.

- **Input:** Accumulated visual and spatial history, initial instruction and the latest navigation instruction appended as a single instruction.
- **Output:** Probability distribution over the next optimal action v_{t+1} (move or stop).

E.3 Question and Answer Generation Model

For both question and answer generation, we adopt the **LANA** model (Wang et al., 2023a) as our baseline module. LANA is a vision-and-language model originally designed to enhance navigation performance by jointly learning to describe past and upcoming paths.

- **Input:** Panoramic observation (X_v) for question generation; or a sequence of panoramic images for the next subpath for answer generation.
- **Output:** Natural language question (q) or guiding answer (a).

In our setup, LANA is initialized with its proposed pretraining and fine-tuned on the RAINbow and RAIN datasets to adapt navigation-style instructions into dialog-style questions and answers.

E.4 Localization Model

For localization, we use a lightweight cross-modal architecture, the Graph Convolutional Network (GCN) Localization (Hahn and Rehg, 2022), as our reproducible baseline. The GCN model formulates localization as a *node-selection* problem. We introduce our graph-based localization, adopting a transformer inspired by **DUET** (Chen et al., 2022b) to improve position inference from dialog.

Repurposing the VLN Architecture. We adapt the DUET architecture to leverage knowledge from VLN pretraining. For localization, we repurpose the architecture by substituting the inputs and reinterpreting the output:

- **VLN Input → Localization Input:** The navigation instruction I is replaced with the **Navigator's question** q, and the partial navigation map $\hat{G}_t$ is replaced with the **entire house graph** G.
- **VLN Output → Localization Output:** The model is tasked to select the most probable node corresponding to the Navigator's current position v_q from all nodes in G: $v_q = \text{DUET}_{\text{LOC}}(G, q)$.

This adaptation, leveraging large-scale pretrained weights, provides robust perceptual grounding and improves localization accuracy.

E.5 Whether-to-Ask

The *whether-to-ask* module governs the interactive nature of the Navigator, deciding when to request external guidance. This module uses a confidence thresholding strategy, where the decision to ask or proceed autonomously is determined by the model's confidence in its immediate next action.

- **Input:** The Navigator's probability distribution over the next possible actions $p(v_{t+1}|\text{context})$.
- **Mechanism:** The most confident action in next possible actions is treated as our candidate action. If the candidate action's probability falls below a fixed threshold τ, the Navigator initiates a question q_j.
- **Output:** Binary decision (Ask=1 or Proceed=0).

F Localization Model on WAY Task

The **Where Are You? (WAY)** (Hahn et al., 2020) is a foundational benchmark for embodied localization, requiring a model to predict its current position v_q within a visual environment (Matterport3D) based solely on a textual query. This task is crucial as it isolates the language grounding ability required by the remote Guide setup in DialNav.

Our model, which integrates the effects of large-scale pretraining from ScaleVLN and the data diversity provided by the RAINbow dataset, has achieved strong performance in the holistic DialNav setup.

We further fine-tune our graph-based transformer localization model on the WAY benchmark to demonstrate that our architecture generalizes well to existing localization tasks. Table 11 provides the results of our model against established single-shot baselines for the WAY localization task, where our model shows superior performance.

Method	Val Seen		Val Unseen	
	Acc@0m ↑	LE ↓	Acc@0m ↑	LE ↓
Human	47.87	6.00	56.13	3.20
Random	0.33	20.80	1.90	18.61
GCN (Hahn and Rehg, 2022)	19.67	10.95	8.64	9.10
LED-Bert (Hahn and Rehg, 2022)	25.57	9.04	21.07	8.82
DiaLoc (Zhang et al., 2024)	25.64	-	7.02	-
CGD (Wang et al., 2025a)	30.21	-	16.72	-
Ours	**32.46**	**6.85**	**23.31**	**7.75**

Table 11: **Localization Performance on the WAY Dataset (Single-Shot).** Comparison of our graph-based transformer model against established single-shot baselines. We report accuracy for exact node prediction (Acc@0m) and Localization Error (LE) in meters across seen and unseen environments.

Model	SR	NSC		DTC	
		All	Succ.	All	Succ.
Val Seen					
Baseline (Han et al., 2025)	30.8	18.65	19.89	3.44	3.57
Ours	58.2	21.30	19.66	4.66	3.49
Val Unseen					
Baseline (Han et al., 2025)	14.5	15.52	14.29	5.61	5.03
Ours	29.0	21.46	20.26	11.08	8.23

Table 12: **Trajectory statistics across success cases.** Comparison of Success Rate (SR), Navigation Step Count (NSC), and Dialog Turn Count(DTC) for all episodes (All) and successful episodes only (Succ.).

G Further Analysis on Table 1

Overall, SPL, NE, and OSR exhibit trends consistent with SR. Separately, NSC and DTC demonstrate that RAINbow and DST encourage exploration in uncertain trajectories. To verify this, we measured NSC and DTC specifically within successful episodes for both the baseline and the models equipped with RAINbow and DST. In the Val Seen split, both scores remain nearly identical to the baseline (DTC: 3.57 → 3.49; NSC: 19.89 → 19.66), indicating that the overall increases in these metrics stem primarily from exploratory behavior in uncertain cases. Building on this, the Val Unseen results show that this uncertainty-driven exploration leads to improved generalization to unseen environments: the models achieve nearly double the success rate while exhibiting increased exploration even within successful episodes (DTC: 5.03 → 8.23; NSC: 14.29 → 20.26).

H Qualitative Results

H.1 Holistic Navigation

Success with Self-Recovery Despite Localization Failure. Figure 14 illustrates a successful DialNav episode that strongly demonstrates the self-recovery capability afforded by our integrated framework. Initially, the Navigator faced a complex environment where the Guide failed to accurately determine the current position. Despite receiving this falsely grounded instruction, the Navigator successfully overrode the erroneous localization estimate and dynamically reconstructed its trajectory (self-recovery) based on goal-relevant information within the dialog.

Failure Due to Instruction-Execution Inconsistency. Conversely, Figure 15 presents a critical failure case that underscores the remaining challenge in seamlessly integrating the conversational policy with the motion execution policy. In this episode, the Guide successfully localized the Navigator and provided a well-grounded answer. However, the Navigator failed to accurately execute the trajectory dictated by the dialogue's instructions, resulting in navigation failure.

H.2 Generated Dialog

We provide a qualitative comparison between dialogs generated by the baseline (trained only on RAIN) and our enhanced model (trained on RAINbow) against the human-annotated Ground Truth (GT) in Table 13. All generated samples were produced under the identical preceding dialog context and current panoramic viewpoint. This analysis clearly demonstrates the improvement afforded by large-scale data augmentation.

The baseline model trained solely on the small RAIN dataset often suffers from redundancy and fragmentation. For instance, in the Answer Example, the RAIN model enters a circular loop ("you have to go to the dining room") and omits essential path information. Similarly, in the Question Example, the generated question is highly repetitive ("a fireplace, a fireplace, and a fireplace"), failing to provide sufficient, diverse visual cues for the Guide to localize the agent. In contrast, the model fine-tuned on **RAINbow** produces significantly more fluent and descriptive instructions. The RAINbow answers include detailed navigational steps grounded in specific objects ("exit the room and turn right to find a staircase") and generally maintain a coherent conversational flow. The RAINbow questions successfully utilize rich visual features ("a wooden table right in the middle," "a large, sliding glass door to the outside") to generate contextually relevant descriptions, which is vital for localization.

However, a qualitative gap still exists compared to the human-written Ground Truth. While RAINbow greatly improves fluency, it sometimes lacks the fine-grained object-level details. Nonetheless, the substantial gain in fluency and coherence achieved by RAINbow validates the effectiveness of our automatic generation pipeline for boosting dialog quality.

I Analysis of Confidence Threshold

We analyze the impact of the Confidence Threshold (τ), the hyperparameter governing the Navigator's decision to ask for help. As the threshold τ increases, the agent triggers a question more often when facing uncertainty. This phenomenon is clearly demonstrated by the Dialog Turn Count (DTC), which shows a consistent increase across both splits as τ approaches 0.9 (Figure 7a).

The increase in dialog frequency directly correlates with the final navigation Success Rate (SR)(Figure 7b). In *Val Seen*, SR shows a sustained increase as the threshold rises. This indicates that dialog is effective at resolving uncertainty and guiding the agent to success in familiar environments. In *Val Unseen*, SR increases up to $\tau \approx 0.5$ and then saturates. This contrast suggests that while dialog provides essential initial information, the model reaches a performance ceiling in novel environments, possibly limited by the quality of answers in complex, unseen settings. The difference in saturation points confirms that the benefit of dynamic dialog is substantially greater in known environments, where the learned policy can best leverage the updated instruction for planning.

Furthermore, we observe that dialog significantly impacts exploration efficiency (Figure 7c). Across both Val Seen and Val Unseen, the Navigation Step Count (NSC) decreases up to $\tau \approx 0.5$ before stabilizing. This demonstrates that proactive dialog is highly effective at reducing redundant exploration, as the agent quickly seeks guidance instead of wandering when uncertain. The stable reduction in NSC followed by stabilization in SR (after $\tau = 0.5$ in Val Unseen) shows a clear correlation between reducing navigational uncertainty via dialog and achieving task success.

J Details on Human-Agent Cooperation

Participants. Ten participants were recruited internally from our research group. All participants signed consent form prior to the evaluation.

Task protocol. Each participant interacted with our agent through a GUI simulator that supported mouse-based navigation in the Matterport3D environment paired with a text-based dialog window. Participants took the role of either Navigator (sending questions and navigating based on the agent's answers) or Guide (receiving the participant Navigator's questions and providing localization-grounded answers), while the agent took the complementary role. The four conditions in Table 6 — Navigator, Guide × Baseline, Ours — were presented in randomized order across both Val Seen and Val Unseen splits. Each participant completed 20–30 episodes in total. After each episode, participants rated the agent counterpart's helpfulness on a 1–5 Likert scale, which we report as Human Score (HS) in Table 6; Success Rate (SR) was computed from task outcomes.

Ethics/IRB. Our human evaluation qualifies for IRB exemption under our institution's policy, as it does not identify individual participants and does not collect or record sensitive personal information as defined by applicable data protection regulations. The evaluation collected only task outcomes (navigation success) and helpfulness ratings; no personal identifiers, demographic information, or sensitive data were recorded.

K Potential Risks

Our work studies dialog-enabled embodied navigation, where language directly affects action decisions in a physical environment. As such systems become more capable, incorrect or misleading dialog may lead to unsafe or inefficient actions. This is particularly important in embodied settings, where execution errors may carry physical consequences.

In addition, our automatic data generation pipeline uses large language and vision–language models to synthesize dialog. Although we validate faithfulness to the intended goal and manually inspect examples, generated data may still contain occasional inaccuracies, biases, or unnatural interaction patterns. If used without careful validation, such artifacts could propagate into trained embodied agents.

We mitigate these risks in two ways. First, our task is evaluated in simulation rather than physical deployment. Second, we validate generated dialog quality and goal faithfulness through both automatic design constraints and human evaluation. Future work should further investigate robustness to adversarial, ambiguous, or misleading dialog in real-world embodied settings.

L Data Safety, Privacy, and Content Checks

We discuss here whether the data used or created in this work contains personally identifying information (PII) or offensive content.

Our dataset is built by repurposing existing VLN resources grounded in Matterport3D indoor environments and by automatically generating navigation dialog from those trajectories. The resulting data consists of navigation paths, scene-grounded object references, and generated question–answer pairs about indoor environments. It is not designed to include names, personal identities, contact information, or other uniquely identifying information about individuals.

M Use of AI Assistants

AI assistants were used in two ways in this work. First, they were used in the research pipeline for automatic data generation, including dialog reformatting and caption-based synthesis, as described in the main paper. Second, they were used during manuscript preparation for writing support, such as grammar correction and improving clarity and flow. All scientific decisions, experimental design, interpretations, and final wording were reviewed and verified by the authors.

Ground Truth (GT)	RAIN (Baseline)	RAINbow (Ours)
Hi I'm on the **top floor**. to my left I see a **wooden desk**, to my right a **wooden cabinet** under a **tapestry**. If I walk past the tapestry there is a **living room area** with a **brown couch, armchair, TV and a chest**.	i 'm in a living room with a fireplace , a fireplace , and a fireplace .	i 'm in this cozy living room with a wooden floor, and there 's a wooden table right in the middle . it 's got a warm vibe with a lamp on the table next to the bed . i can see a white chair with a purple seat in front of the window .
Hi, I am at a **gym place** full of **different gym equipments**. If there isn't any other gym room in the house, it should be easy to find out my current location. Otherwise, I will find more distinguishable features of the room. Thank you for helping me find the goal room.	i 'm in a gym room with a lot of . there are two black chairs and a water dispenser .	i 'm in a big , open gym with a bunch of exercise machines. there 's a treadmill on the left and a bench in the middle . i also see a large , sliding glass door to the outside .

(a) Generated Question

Ground Truth (GT)	RAIN (Baseline)	RAINbow (Ours)
Come out of the **bathroom** and you'll see a big **bedroom** with a **big bed** and lots of **sofas**. The bed has white sheets and **blue pillows**. Come out of the bedroom through the door that's left to the **Tv**. Next to the door is a **picture of a sailboat**. Once you've reached the hallway that has some writings hanging on the wall to your left, let me know.	ok , i found you . please walk out of the bathroom and go straight until you see a bedroom on your left . turn right and go straight until you see a bedroom on your left . turn right and go straight until you see a bedroom on your left . go up the stairs and let me know	alright , head out of the bathroom through the open arch and walk straight to the bedroom entrance . go through the doors , turn left to find the lounge area in the center . look slightly right to spot a painting of a ship on the left wall . then , go through the open door in front of you and walk toward the photo frame on the front wall . turn right and head straight to the staircase on your left , climbing up .
Okay, go to the **dining table** and walk through the **right side** of the dining table. You'll see a room both in left and right. **Enter the right one**. Then, you'll see a **stair on your right side**. Can you see this stair?	ok , you 're in the dining room . you have to go to the dining room .	no problem ! from the dining table , turn around and head towards the wall . slightly turn left and go to the open door . exit the room and turn right to find a staircase. go up the steps , then turn right again to see another open door . enter that room and stand in front of the mirror . you 've made it to the bathroom !

(b) Generated Answer

Table 13: **Qualitative Comparison of Generated Dialog.** Comparison of Ground Truth (GT) questions and answers versus models trained on RAIN and RAINbow data. Bold text indicates essential content in the Ground Truth (GT), blue text denotes overlapping content found in the generated outputs, and red text highlights non-fluent content. RAINbow significantly improves fluency and reduces redundancy compared to the RAIN baseline.

Q: Hi. I'm bedroom, which has one bed. On the bed there is two green pillows. And also, there is a big picture, and I can't guess what is it about. It's very simple room.

A: there are 7 bedrooms in this house. I think you're on the 1st floor. If I'm correct, you'll see a very long corridor outside of the room. Can you see?

Q: Yes, I can see very long hallway, and also there is a chair in the middle of the hallway.

A: okay, at the end of the corridor, you'll find a staircase on your right. Go all the way up. I mean, there'll be another half story when you think you're at the top, so look carefully :) At the real end of the staircase, you'll find a kitchen. That's your goal room.

Q: I think I'm at that kitchen, there is a statue of woman in front of the long wooden table. Am I at the goal room?

A: Yes correct. good work

(a) RAIN

Q: I'm in a small kitchen with a wooden door. I see a painting on the wall, it's a big portrait of a guy in a blue jacket, looks like Jimi Hendrix holding a guitar.

A: Got it! Head downstairs and take a left. Just stop in front of the doorway.

Q: Now I'm in a cozy little bedroom with stairs going up. It's quite narrow, and there's a guitar laying on the floor.

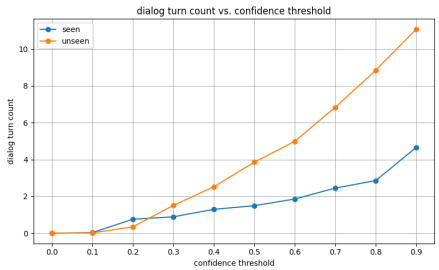
A: Alright, walk right past the stairs and pause in the entryway at the end of the hall.

Q: I'm in a spacious bedroom with a bright white interior, and there's a wooden door with a lock.

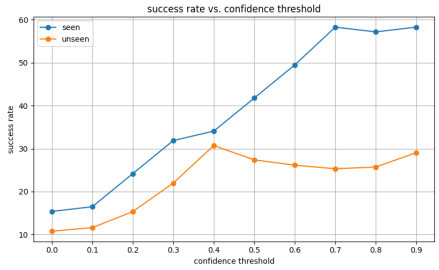
A: Perfect! Exit the bedroom, go down the stairs, and head through the doorway next to the plant with white flowers. Stop beside the painting of the man with the guitar. You've made it to the hallway.

(b) RAINbow

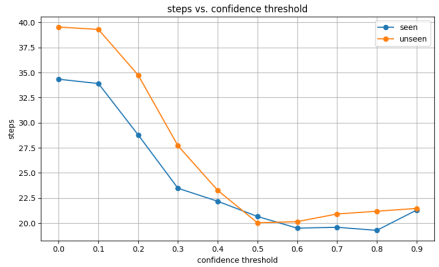
Figure 6: **Qualitative comparison of RAIN (Human) and RAINbow (Generated) dialogs.** The figure presents examples of complex, multi-turn dialogs. The RAIN example (top) illustrates the necessity of detailed scene descriptions for successful remote localization. The RAINbow example (bottom) demonstrates that our automatic generation pipeline successfully replicates this essential dialog structure.



(a) Dialog Turn Count vs. Confidence Threshold

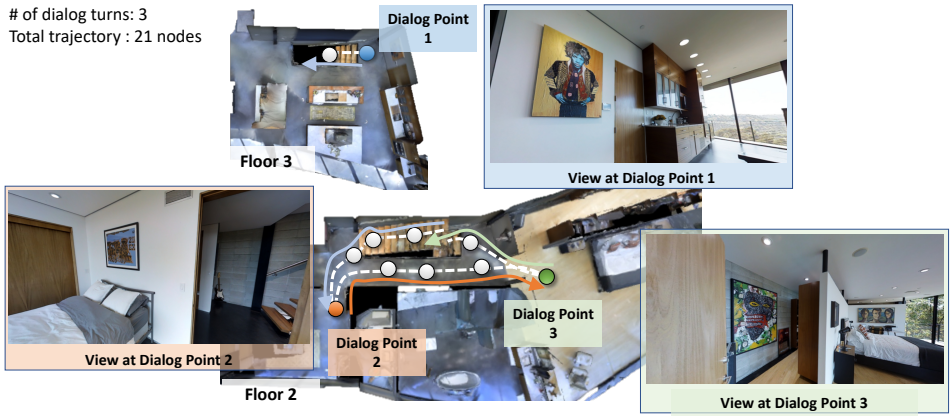


(b) Success Rate vs. Confidence Threshold



(c) Navigation Steps vs. Confidence Threshold

Figure 7: **Impact of Confidence Threshold (τ) on performance.** Adjusting the threshold value allows the Navigator to balance autonomous navigation with dialog requests, showing the trade-off between task success and exploration efficiency.



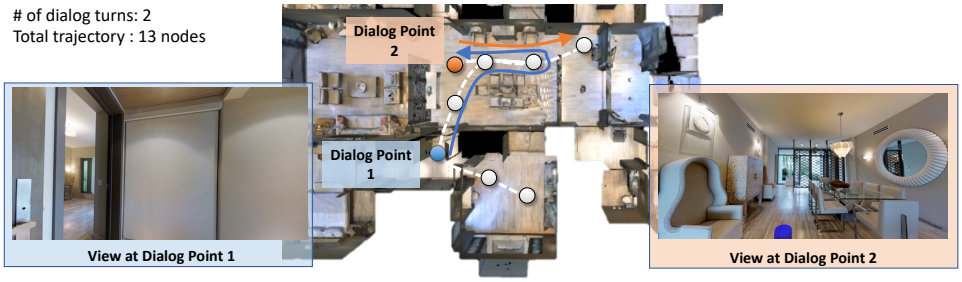
Initial Instruction: The target room has *painting*.

At Dialog Point 1
Q ( ): I'm in a small kitchen with a wooden door. I see a painting on the wall, it's a big portrait of a guy in a blue jacket, looks like Jimi Hendrix holding a guitar.
A ( ): Got it! Head downstairs and take a left. Just stop in front of the doorway.

At Dialog Point 2
Q ( ): Now I'm in a cozy little bedroom with stairs going up. It's quite narrow, and there's a guitar laying on the floor.
A ( ): Alright, walk right past the stairs and pause in the entryway at the end of the hall.

At Dialog Point 3
Q ( ): I'm in a spacious bedroom with a bright white interior, and there's a wooden door with a lock.
A ( ): Perfect! Exit the bedroom, go down the stairs, and head through the doorway next to the plant with white flowers. Stop beside the painting of the man with the guitar. You've made it to the hallway.

(a)



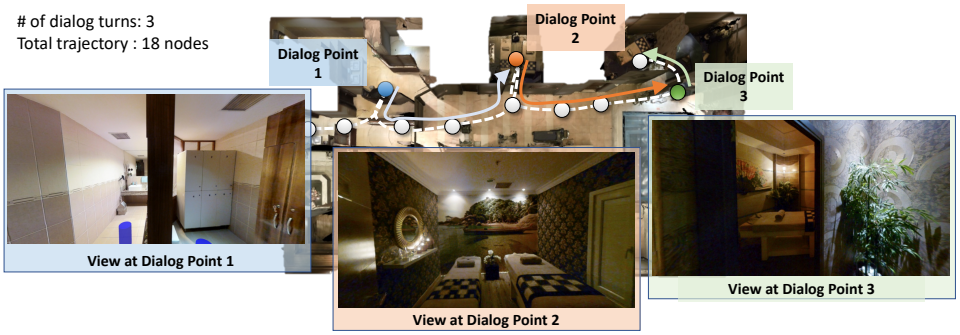
Initial Instruction: The target room has *chair*.

At Dialog Point 1
Q ( ): I'm in a long, narrow hallway with white walls that are covered in fabric, giving it a clean, modern vibe.
A ( ): Got it! Turn left, then right, and make your way to the left side of the dining room table.

At Dialog Point 2
Q ( ): I followed those directions and now I'm in a spacious dining room that has a large table with chairs, a couch, and a bench. It feels really open with the living room attached, and there are some decorative items like a vase and a potted plant around.
A ( ): Nice! Walk past the glass table on your right, head toward the lattice wall, and just wait at the opening there. You've arrived at your destination.

(b)

Figure 8: Examples of RAINbow data (1)



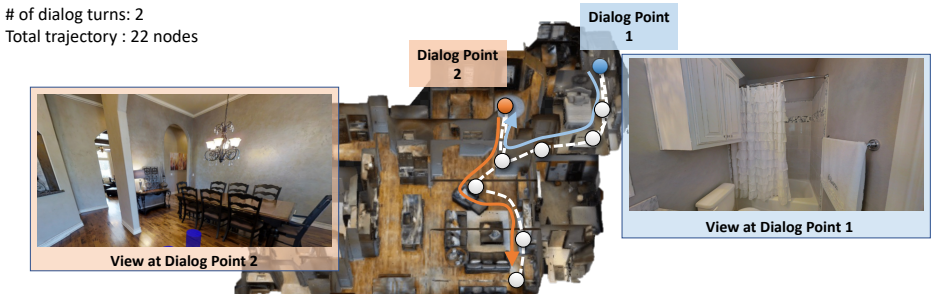
Initial Instruction: The target room has *lamp*.

At Dialog Point 1
Q ( ): I'm in a small, narrow utility room with an open doorway. It's mostly white, with tiled floors and walls. It feels more like a workspace or storage area.
A ( ): Got it. Step out of the bathroom, head into the hall, and make a left. Keep going straight, and you'll find the massage room as the second door on your left. You'll be right between the two beds.

At Dialog Point 2
Q ( ): I made it to a spa room! There's a big painting of a tropical island on the wall, and everything feels really relaxing. There are two couches, one on the left and one on the right, plus a chair in the middle.
A ( ): Nice! Now, exit the massage room and take a left. Walk toward the pool, then head through the door to the left of the pool. Stop near the plant in the hallway.

At Dialog Point 3
Q ( ): I'm now in a dimly lit hallway, with a bathroom door right in front of me. It's got a bit of a mysterious vibe.
A ( ): Enter that room through the door, then turn right and walk towards the bed. You've reached the spa/sauna.

(a)



Initial Instruction: The target room has *rug*.

At Dialog Point 1
Q ( ): I'm in a bathroom with a clean vibe, featuring a white shower curtain and a towel on the wall.
A ( ): Great! Exit the bathroom through the open door on the left of the curtain. Then, turn slightly right to leave the bedroom through another open door. Move forward to the archway and step into the left side, where you'll see the sink and four chairs. Turn right and head towards the circular rug on the floor. Go past the rug and head through the glass door. Keep moving forward, you'll spot a sofa chair on your right and a table behind you.

At Dialog Point 2
Q ( ): I followed your directions and ended up in a spacious dining room. There's a wooden floor, with a dining table surrounded by chairs and a cozy rug underneath. This room seems to connect to the living room.
A ( ): Perfect! From the opening, turn around to face the dining area. Next, turn slightly left to find an archway and walk past it. After that, turn a bit left again to see the living room ahead, with a counter and stools on your right. Walk towards the second stool, then turn left to reach the coffee table in front of you. Continue towards the right side of the fireplace where you'll notice a black sofa with a blue pillow on your left and a dresser on your right with some decor. Finally, walk toward the window where a patio table with chairs is visible. You've reached the living room!

(b)

Figure 9: Examples of RAINbow data (2)

Prompt for Simple Panoramic Caption

```
Generate a sentence that describes the given panoramic image. Start your question with "I'm in ..." or "I can see ...".
```

(a) Prompt for Simple Panoramic Caption

Prompt for Region Grounded Caption

```
Please describe the {region} in this image.
```

(b) Prompt for region grounded caption

Prompt for Object Caption

```
Please describe the {obj} in this image. If there's no {obj}, response 'None'
```

(c) Prompt for object caption

Prompt for Region and Object Grounded Caption

```
You will be provided with a panoramic image caption describing an entire space and detailed captions of key objects visible in that image. Based on the input, describe the surrounding environment clearly and concisely. Your description should include the structure, layout, and the locations of key objects. Do not add anything, but just the sentence describing around.
```

(d) Prompt for region and object grounded caption

Figure 10: Prompts used to generation captions for navigation questions.

Simple caption	*"I'm in a room with a black counter and a large marble wall."*
Region grounded caption	*"The dining room in this image features a large, modern design with a black and white color scheme. The room is adorned with a marble wall, which adds a touch of elegance to the space. There is a large window that allows natural light to fill the room, creating a bright and inviting atmosphere. A potted plant is placed in the room, adding a touch of greenery and life to the space. Additionally, there is a chair in the room, providing a comfortable seating option for those who wish to dine or relax in the dining area"*
Object caption	*The dining table in this image is a black table with a vase on it., The pillow is white., The couch in this image is a large, gray sectional sofa., The chair in the image is a modern, sleek design with a gray color. It is positioned at a black dining table., The chair in the image is a modern, minimalist design with a white seat and black legs. It is positioned at a black dining table.*
Region and object grounded caption	
1-shot, 50% objects	*"I'm in a modern dining room with a black and white color scheme, surrounded by a marble wall. There's a large window letting in plenty of natural light, and a potted plant adding some greenery to the space."*
1-shot, 70% objects	*"I'm standing in a modern dining room with a black and white color scheme, surrounded by a marble wall. The room is filled with natural light from a large window and features a potted plant and a chair for seating.."*
3-shot, 50% objects	*"I'm in a modern dining room with a black and white color scheme, surrounded by a marble wall. There's a large window letting in plenty of natural light, and a potted plant adds a touch of greenery to the space."*
3-shot, 70% objects	*"I'm in a modern dining room with a black and white color scheme, featuring a marble wall and a large window that lets in plenty of natural light. There's a comfortable chair and a potted plant adding a touch of greenery to the space."*
5-shot, 50% objects	*"I'm standing in a modern dining room with a black and white color scheme, surrounded by a marble wall and a large window that lets in plenty of natural light. A potted plant adds a touch of greenery to the space, and a sleek gray chair sits at a black dining table.*
5-shot, 70% objects	*"I'm in a modern dining room with a black and white color scheme. The room features a marble wall, a large window, and a potted plant, creating a bright and inviting atmosphere. There's a chair at a black dining table, providing a comfortable seating option."*

Figure 11: Example of generated captions from three different captioning schemes.

Prompt for Goal-conditioned Reformat

```
You will be provided with a sequence of sentences describing a path.
Your task is to rewrite the path description in two versions:
1. with_goal: A version that includes reaching the final destination
2. without_goal: A version that excludes any mention of reaching the destination
Both versions should:
- Be concise while preserving key spatial references
- Focus on objects, rooms, and layout
- Use directions only when necessary
- Remove terms implying current situation (Start facing  , You're in a  , You're facing)
For with_goal version, Always mention that you've arrived at the destination at the end of the description with expressions
like: That's the goal, You have arrived, That's your destination, The goal, You've reached your destination, This is the
final spot, You're at the destination, Mission accomplished, End point reached, You've made it.

For without_goal version, Remove any mentions of arriving at or reaching the final destination.
```

(a) Prompt template for goal-conditioned path reformatting.

Prompt for Natural Dialog Reformat

```
You will rewrite a dialog between a Navigator and a Guide in an indoor navigation task.
TASK CONTEXT:
- The Navigator and Guide work together to reach a goal room.
- Both Navigator and Guide know about a shared object in the goal room.
- The Guide knows the goal room but the Navigator does not.
- The object may appear in other rooms, so the Navigator must ask clarifying questions.
- The Guide cannot see the Navigator's position but knows the environment layout.
- Dialog alternates turns: Navigator → Guide → Navigator . . .  (no limit on turns).
- Dialog always starts with the Navigator.
OBJECTIVE:
- Rephrase the dialog to sound like natural conversation.
- Style:
* Use casual, friendly language ("Ok", "Alright", "Got it").
* Add acknowledgments ("I see", "Understood").
* Smooth transitions between turns.
* Vary vocabulary, keep meaning the same.
- Actively incorporate mentions of the shared object in the conversation.
- The Guide's final response should mention reaching the goal room.
RULES:
1. Keep the same number of turns and same speakers (Navigator/Guide).
2. Do not add or remove turns.
3. Do not add new objects, rooms, or details not in input.
4. Never use phrases like "You're looking at. . . " or "You are facing. . . ". - Instead: ask questions ("Do you see. . . ?")
or omit.
5. Do not drop or shorten navigation instructions.
6. Preserve all key details:
- Objects and attributes (color, shape, material).
- Room names and types.
- Spatial relations and directions.
7. Do not use em dashes (—). Use commas, periods, or "and".
OUTPUT FORMAT:
Return ONLY valid JSON:
"reformatted": [
"Navigator": "...", "Guide": "..." ,
"Navigator": "...", "Guide": "..."
]

No extra text, no markdown, no explanations.
```

(b) Prompt template for natural dialog reformatting.

Figure 12: Prompt for dialog reformatting.

Example
Original Instruction
"We start off looking at the corner of a wall. As you turn around you see a large open are with doorways to many rooms around you. On one of the walls you see a white rectangular box. Take a step towards that white rectangular shape and turn to look down the hall and you will see two chairs directly in front of you down the hall. And before those two chairs on the right there is an open walkway. Take a step towards that open walkway on your right. You will be looking into a room that has brown cabinets and a shelf with a white pillow and radio looking thing on it and you are done."
Reformatted (without goal)
"From the corner of a wall, turn to see a large open area with doorways to various rooms. Approach the white rectangular box on the wall, then look down the hall to see two chairs ahead. To the right of the chairs, there's an open walkway. Step into the walkway to view a room with brown cabinets and a shelf containing a white pillow and a radio."
Reformatted (with goal)
"Turn away from the wall to find a large open area with doorways to various rooms. Move toward the white rectangular box on the wall, then look down the hall to see two chairs ahead. To the right of those chairs is an open walkway; step into it. You'll see a room with brown cabinets and a shelf containing a white pillow and a radio. You've reached your destination."

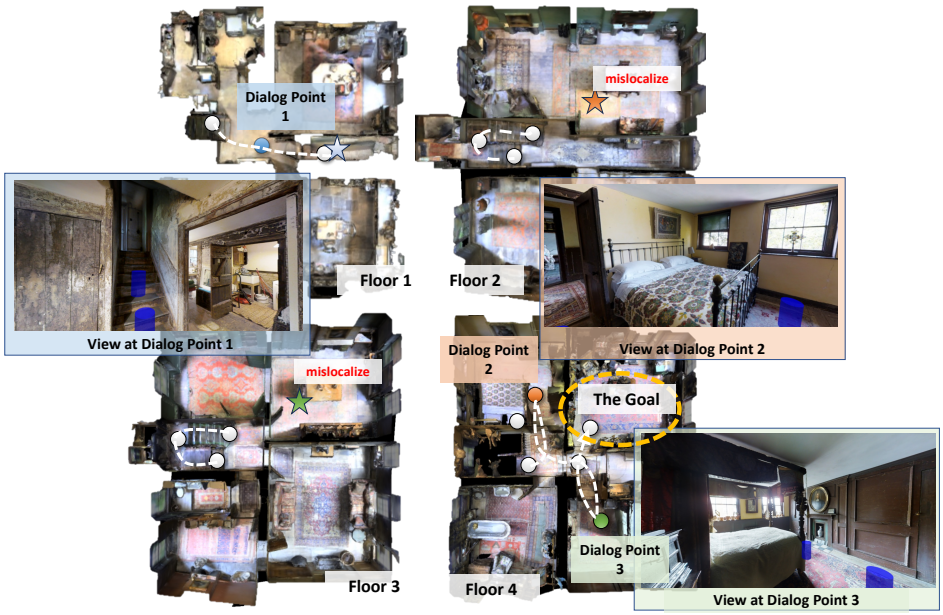
(a) Instruction Refining.

	Before Reformat	After Reformat
Q1	I'm in a long hallway filled with books from floor to ceiling. The bookshelf stretches the entire length of the hallway, packed with books of all sizes and types, creating an impressive display of knowledge and resources.	I'm in this long hallway packed with books all the way up to the ceiling. The bookshelf runs the entire length of the hallway, filled with all sorts of books, creating an amazing sight.
A1	You are standing in a library, to the right you have a book shelf, turn back from the place where you are standing, walk straight, till the end, turn right, walk straight, until you reach the book shelf in-front of you, turn right, walk straight, walk straight until you reach the book shelf in-front of you and a staircase on your right, get down the stairs completely. Slightly turn right, in-front of you there is a large table with chairs around it, to the right you have a book shelf, stand between the table and book shelf, turn back. Sorry, turn right, in-front of you there is an open door walk to the open door, in-front of you there is a staircase, stand on the staircase, that's your end point.	Awesome! In the library, head straight to the end and take a right. Keep going until you reach another bookshelf, then make another right and head toward the bookshelf with a staircase on its right. Go down the staircase completely, and then turn a bit right to find a big table surrounded by chairs and another bookshelf on the right—stand right between those two. Next, turn right toward the open door and walk through it to reach the staircase.
Q2	The library in this image is a room with a stone archway and a white wall. The room appears to be a part of a university, as it is described as a "college library." The stone archway adds a unique architectural element to the space, making it visually appealing. The white wall complements the stone archway, creating a clean and elegant atmosphere. The room is likely used for studying, research, and quiet contemplation, as it is a library setting.	I see the library has a stone archway and a white wall. It feels like a university space, really charming.
A2	You are now facing the white color door, from there take a left turn. Start going down the stairs and stand on the tenth step, you will find your destination point.	Perfect! Now, turn left from the white door and go down the stairs until you hit the tenth step. That's your goal.

	Before Reformat	After Reformat
Q1	I am in a large, wooden wardrobe with a mirrored door. The closet is filled with various clothing items, including shirts and suits, and is well-organized, making it easy to find and access the clothes stored inside.	I'm inside this large wooden wardrobe with a mirrored door. It's packed with all kinds of clothes like shirts and suits, and everything is organized really well, so it's easy to find what I need.
A1	Walk down the hallway to your right and move past the bed. Walk through the door on the right of the painting of thr girl and wait there.	Alright! Start by heading down the hallway to your right and walk past the bed. Then, go through the door on the right side of the girl's painting and just wait there.
Q2	I'm in a curved hallway with a staircase and glass railing, surrounded by various artworks on the walls, including paintings and photographs.	I've made it to a curved hallway that has a staircase and a glass railing. The walls are decorated with various artworks, including paintings and photos.
A2	down more stairs, and into the little room left of the silver mirrored planter	Nice! Now, head down those stairs and enter the small room to the left of the silver mirrored planter. You've finally arrived at your destination.

(b) Conversational Smoothing (Q: Question, A: Answer).

Figure 13: **Example of dialog reformatting.** (a) Examples of **Step 1 (Instruction Refining)**, where raw VLN instructions are reformatted into 'without goal' and 'with goal'. (b) Examples of **Step 2 (Conversational Smoothing)**, where the sequence of captions and refined instructions is paraphrased into fluent dialog.



Initial Instruction: The target room has *frame*.

At Dialog Point 1

Q ( ): i 'm in this hallway with a concrete wall that has a rustic vibe . there 's a staircase leading up , but no furniture around .
A ( ): okay . come out of the room and you 'll see the stairs going up . go up the stairs and at the end of the stairs , you will see a room with a stairway on your left . go into that room and that is the goal room .

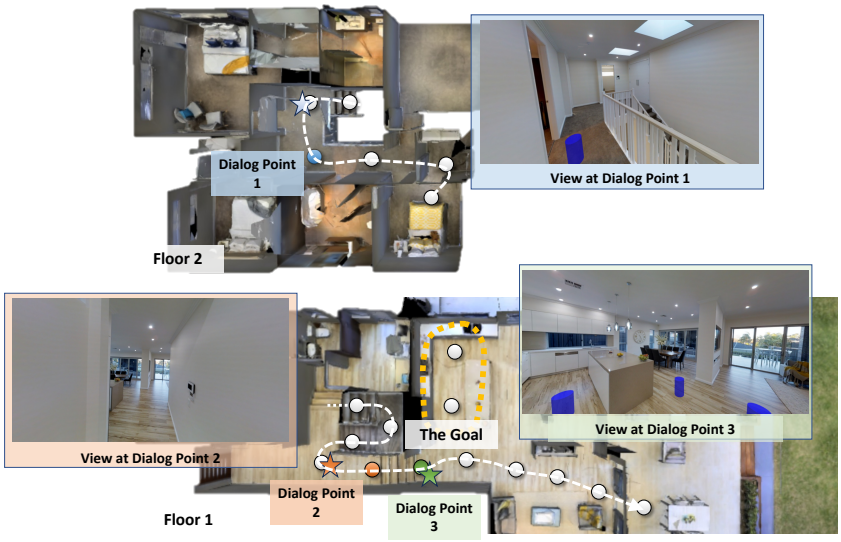
At Dialog Point 2

Q ( ): I'm in a hallway with a wooden door on my right . there 's a wooden dresser with a lamp on it , and a nearby window , plus a wooden dresser and a blue and white striped armchair .
A (, mislocalize **):** come out through the door , come up the stairs . turn left into a lounge . goal .

At Dialog Point 3

Q ( ): i 'm in a hallway with a wooden door on my left and a window on my right . the door is open , and i can see a room with a purple and gold armchair and a wooden cabinet .
A (, mislocalize **):** alright , turn around and head towards the open door on your left . go through that door and you 'll find a staircase on your right . climb up the stairs , then turn left to enter the next room . move forward until you 're in front of the bookshelf . you 've made it to the goal room .

Figure 14: **Successful example of DialNav episode of our Navigator and Guide agent in Val Unseen environment.** Although the agent failed to localize its current position in the complex environment, it successfully acquired crucial information about the target room through dialogue. This allowed the agent to self-recover its trajectory and successfully navigate to the destination.



Initial Instruction: The target room has *plant*.

At Dialog Point 1
Q ( ): i 'm in a long , narrow hallway with a white color scheme . there 's a staircase going up to the upper floor , and it 's well - lit with a nice painting on the wall .
A ( ): perfect ! go down the stairs and take a left . walk past the kitchen , then take a left . you 'll find the pantry there .

At Dialog Point 2
Q ( ): i 'm in a long hallway with a white wall and a wooden floor . there 's a small , thin white marble vent on the wall .
A ( ): perfect ! go down the stairs and take a left . walk past the kitchen , then take a left . you 'll find the pantry there .

At Dialog Point 3
Q ( ): i 'm in a modern kitchen with a big open layout . there 's a staircase leading up , and i see a sink , a fridge , and a dining table with chairs .
A ( ): great ! turn around and head towards the refrigerator . then , slightly turn left and walk towards the stove . after that , turn left again and go towards the stove .

Figure 15: **Failure case of DialNav episode of our Navigator and Guide agent in Val Unseen environment.** Despite achieving successful localization and receiving a well-grounded answer regarding the path, the navigation agent failed to follow the instruction derived from the dialogue, resulting in a failure to reach the destination.