DOROTHIE: Spoken Dialogue for Handling Unexpected Situations in Interactive Autonomous Driving Agents

Ziqiao Ma†, Ben VanDerPloeg†, Cristian-Paul Bara†, Yidong Huang†, Eui-In Kim†, Felix Gervits‡, Matthew Marge‡, Joyce Chai‡

†University of Michigan     ‡U.S. Army Research Laboratory
{marstin,bensvdp,cpbara,owenhji,euiink,chaijy}@umich.edu
{felix.gervits,mattthew.r.marge}.civ@army.mil

Abstract

In the real world, autonomous driving agents navigate in highly dynamic environments full of unexpected situations where pre-trained models are unreliable. In these situations, what is immediately available to vehicles is often only human operators. Empowering autonomous driving agents with the ability to navigate in a continuous and dynamic environment and to communicate with humans through sensorimotor-grounded dialogue becomes critical. To this end, we introduce Dialogue On the ROad To Handle Irregular Events (DOROTHIE), a novel interactive simulation platform that enables the creation of unexpected situations on the fly to support empirical studies on situated communication with autonomous driving agents. Based on this platform, we created the Situated Dialogue Navigation (SDN), a navigation benchmark of 183 trials with a total of 8415 utterances, around 18.7 hours of control streams and 2.9 hours of trimmed audio. SDN is developed to evaluate the agent’s ability to predict dialogue moves from humans as well as generate its own dialogue moves and physical navigation actions. We further developed a transformer-based baseline model for these SDN tasks. Our empirical results indicate that language-guided-navigation in a highly dynamic environment is an extremely difficult task for end-to-end models. These results will provide insight towards future work on robust autonomous driving agents.

1 Introduction

In embodied agents such as autonomous vehicles (AVs), highly dynamic environments often lead to unexpected situations, such as challenging environment conditions (e.g., caused by weather, light, obstacles, etc.), influence of other agents, and change of the original goals. In these situations, the agent’s pre-trained models or existing knowledge may not be adequate or reliable to make a corresponding decision. What is immediately available to help the agent is often only human partners (Ramachandran et al., 2013). As they are not programmers who can readily change the code in the field, approaches that enable natural communication and collaboration between humans and autonomy become critical (Spiliotopoulos et al., 2001; Weng et al., 2016). Although recent years have seen an increasing amount of work in natural language communication with robots, and especially the many benchmarks that have been developed for navigation by instruction following (Roh et al., 2020; Vasudevan et al., 2021; Shridhar et al., 2020; Padmakumar et al., 2022), little work has been done to study language communication under unexpected situations, particularly in the context of AVs.

To address this limitation, we have developed Dialogue On the ROad To Handle Irregular Events (DOROTHIE), an interactive simulation platform built upon the CARLA simulator (Dosovitskiy et al., 2017) to specifically target unexpected situations. The DOROTHIE simulator supports Wizard-of-Oz (WoZ) studies through a novel duo-wizard setup: a collaborative wizard (Co-Wizard) that collaborates with the human to accomplish the tasks, and an adversarial wizard (Ad-Wizard) that generates unexpected situations (e.g., creating road obstacles, changing weather conditions, adding/changing goals, etc.) on the fly. Using DOROTHIE, we collected the Situated Dialogue Navigation (SDN) dataset of 183 trials between a Co-Wizard and human subjects to collaboratively resolve unexpected situations and complete navigation tasks through spoken dialogue.

The SDN dataset contains multi-faceted and time-synchronized information (e.g., first-person view of the environment, speech input from the human, discrete actions, continuous trajectory and control sig-
Table 1: Comparison of language-conditioned task completion settings in terms of Environment Fidelity (Simulated, Panoramic, Physical), Environment Continuity (Discrete, Continuous), Communication Form (Freeform Dialogue, Restricted Dialogue, Multiple Instructions), Language Granularity (High: Goal, Low: Step/Movement), Control Granularity (High: Action, Low: Control), Demonstration Collection (Human, Planner), Modalities (Language, Vision, Map, Speech), Instruction Type (Replanning, Adaptation, Navigation, Manipulation), Action Space (Discrete, Continuous).

2 Related Work

Our work is mostly related to language-conditioned navigation tasks (Anderson et al., 1991; MacMahon et al., 2006; Paz-Arangelo and Tsarfaty, 2019) and particularly recent work on embodied agents that learn to navigate by following language instructions (Gu et al., 2022). Table 1 summarizes the comparison between our work and previous work. Below we highlight some key differences.

Replanning in Unexpected Situations. Most simulated environments assume that only the task agent can change the state of the world through navigation and/or manipulation. In outdoor settings, the agent operates in a highly dynamic environment where unexpected changes to the world can often occur due to, e.g., walking pedestrians, moving vehicles, lighting, and weather conditions. While previous studies have explored misleading (Roh et al., 2020) or perturbed (Lin et al., 2021) instructions, no prior work has looked into how language instructions can help agents adapt in these unexpected situations. To our knowledge, SDN is the first dataset where language is used to assist agents to replan their goals, paths, and trajectories.

Free-Form Communication. Most prior work adopts either simple instruction-following (Chen et al., 2019; Shridhar et al., 2020; Vasudevan et al., 2021), or restricted QA dialogue (Chai et al., 2018; Thomason et al., 2020; Gao et al., 2022) that only allows the agent to ask for help. Except for some recent work in human-robot dialogue (She and Chai, 2017; De Vries et al., 2018; Banerjee et al., 2020; Padmakumar et al., 2022), few efforts have supported fully free-form communication where agents can ask, propose, explain, and negotiate under ambiguity or confusion. To the best of our knowledge, SDN is the first benchmark to enable navigation in autonomous driving agents conditioned on free-form spoken dialogue.
Continuous Navigation. In discrete navigation, agents take discrete actions, e.g., tele-transport in a pre-defined grid world (De Vries et al., 2018) or a navigation graph with sparsely sampled panoramas at each node (Chen et al., 2019; Vasudevan et al., 2021). More recently, researchers proposed a continuous navigation setting (Krantz et al., 2020; Hong et al., 2022) by converting discrete paths on navigation graphs into trajectories. Unfortunately, these agents are still limited with a discrete action space such as forward 0.25m. This becomes unnatural in outdoor settings because the default behaviour of outdoor driving agents (e.g., autonomous vehicles) is lane-following instead of staying still. We instead follow the settings of mobile robot navigation (Roh et al., 2020; Irshad et al., 2021), where the agents are controlled by a continuous action space with physics like throttle and steering, leading to continuous control signals with long-range trajectories.

3 Dialogue On the ROad To Handle Irregular Events (DOROTHIE) Simulator

Motivated by the wide availability of software simulations for autonomous vehicles (Rosique et al., 2019), we set up our experiment in CARLA (Dosovitskiy et al., 2017), a driving simulator for autonomous vehicles. We developed a novel framework, Dialogue On the ROad To Handle Irregular Events (DOROTHIE) (as shown in Figure 1), to study situated communication under unexpected situations based on the Wizard-of-Oz (WoZ) paradigm (Riek, 2012; Kawaguchi et al., 2004; Hansen et al., 2005; Eric et al., 2017). In WoZ, a human participant is typically instructed to interact with an autonomous agent to complete a set of tasks. The agent’s behaviors, however, are controlled by a human “wizard” (i.e., a researcher).

One important novelty of our framework is that it extends the traditional WoZ approach by introducing a pair of wizards. In our duo-wizard setup, a Co-Wizard controls the agent’s behaviors and carries out language communication with the human participant to jointly achieve a goal, and an Ad-Wizard creates unexpected situations on the fly. The Co-Wizard and the participant need to resolve these unexpected situations as they arise.

3.1 Interface for Co-Wizard Activities

We found in pilot studies that a low-level, free-form controller is not desirable due to the poor quality of demonstrated trajectories and high cognitive load...
on the Co-Wizard. In line with prior work (Roh et al., 2020; Codevilla et al., 2018; Mueller et al., 2018), we developed a set of high-level physical actions from pilot studies for the Co-Wizard to control the vehicle. Each action is mapped to a rule-based local trajectory planner to generate a list of waypoints that the vehicle will drive through. The continuous control (steering, throttle, brake) of the vehicle is performed by a PID controller.

In a complex navigation task with multiple subgoals, belief tracking over plans, goals, task status, and knowledge becomes crucial (Ma et al., 2012; Misu et al., 2014). Besides controlling the vehicle and communicating with the participant, the Co-Wizard also annotates the intended actions (referred to as mental actions) during and after the interaction, e.g., by noting down the navigation plan by clicking junctions on the intended trajectory from current position to the destination. The set of the physical and mental actions is described in Figure 2 and more implementation details are available in Appendix A.6.

<table>
<thead>
<tr>
<th>Physical Actions</th>
<th>Args</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>LaneFollow</td>
<td></td>
<td>Default behaviour, follow the current lane.</td>
</tr>
<tr>
<td>LaneSwitch</td>
<td>Angle (Rotation)</td>
<td>Switch to a neighboring lane.</td>
</tr>
<tr>
<td>UTurn</td>
<td>Angle (Rotation)</td>
<td>Turn to a connecting road at a junction.</td>
</tr>
<tr>
<td>Stop</td>
<td></td>
<td>Make a U-turn to the opposite direction.</td>
</tr>
<tr>
<td>Start</td>
<td></td>
<td>Start the vehicle manually.</td>
</tr>
<tr>
<td>SpeedChange</td>
<td>Speed (km/h)</td>
<td>Change the desired speed by 5 km/h.</td>
</tr>
<tr>
<td>LightChange</td>
<td>Light State (on/off)</td>
<td>Change the front light state.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mental Actions</th>
<th>Args</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>PlanUpdate</td>
<td>Landmark, Status</td>
<td>Indicate intended trajectory towards a destination.</td>
</tr>
<tr>
<td>GoalUpdate</td>
<td>Landmark</td>
<td>Indicate current goal as an intended landmark.</td>
</tr>
<tr>
<td>StatusUpdate</td>
<td>Landmark, Status</td>
<td>Indicate a change in task status.</td>
</tr>
<tr>
<td>KnowledgeUpdate</td>
<td>x,y</td>
<td>Guess the location of an unknown landmark.</td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>Other belief state updates.</td>
</tr>
</tbody>
</table>

Table 2: The space of primitive physical actions and mental actions of the Co-Wizard.

### 3.2 Interface for Ad-Wizard Activities

The Ad-Wizard is able to introduce environmental exceptions and task exceptions.

- **Environmental Exceptions**: Triggered by changes to the environment. These include direct environmental changes, which challenge the vehicle’s perceptual processing and motivate participants to request for adaptations without changing the plan or goal (e.g., drive slowly in foggy weather and turn the headlights on at night). Environmental exceptions can also be introduced by creating roadblocks, which motivate new plans by blocking the original ones.

- **Task Exceptions**: Brought by changing the tasks specified in the storyboard by deleting, adding, or changing a landmark to visit. The Ad-Wizard will send a message to prompt the participant in the message interface with appropriate context, and modify the task interface that specifies the landmarks to visit. Since the Co-Wizard does not have a task interface, the participant needs to communicate with the Co-Wizard in natural language to inform the status of a subgoal, especially when a change of current subgoal is indicated by the Ad-Wizard.

The rich dynamics of the environment and tasks in DOROTHIIE create uncertainty and ambiguity, which requires the Co-Wizard to actively initiate conversation with the human partner and find a way to handle these unexpected situations collaboratively. More illustrated details of the Ad-Wizard interface is available in See Figure 10 in Appendix A.7.

### 3.3 Data Collection

Using DOROTHIIE, we recruited 40 naïve human subjects as participants for data collection. Each subject went through an average of 4.5 sessions. In each session, a storyboard was given to the subject which required the agent to visit two to six landmarks/destinations. Each storyboard was generated from four different towns, with all task templates, landmark locations, street names and departure locations randomly shuffled. While shown the map, the Co-Wizard (an experimenter) did not have access to some of the destinations, e.g., the location of a friend’s house or a person to pick up. Such knowledge disparities motivate rich situated communication and challenge the agent to understand language instructions of different granularity. As the Co-Wizard and the human subject communicated with each other to achieve the goal, the Ad-Wizard (another experimenter) was tasked to create different types of unexpected events that were relevant to the current goal. The knowledge disparity and unexpected events together drive the communication. Details of the task setups are available in Appendix A.4.

### 4 Situated Dialogue Navigation (SDN)

Our data collection effort has led to the Situated Dialogue Navigation (SDN), a fine-grained outdoor navigation benchmark. Each session was replayed at 10 FPS following prior work (Roh et al., 2020) to obtain multi-faceted and time-synchronized information, e.g., a first-person view of the environment, speech input from the participant, discrete actions,
Dialogue Moves

![Diagram of Dialogue Moves]

Figure 2: The coding scheme of dialogue moves as a decision tree. The leaf nodes of the decision tree specify the set of dialogue moves we used for annotation.

4.2 Data Statistics

The dataset is split into training, validation, and test sets and defines seen (Town 1, 3, 5) and unseen (Town 2) sub-folds for validation and test. The SDN dataset captures rich dialogue behaviors between the human and the agent to collaboratively resolve unexpected situations and achieve joint goals. Table 3a shows some basic statistics.

(a) Dataset Statistics and split information.

(b) The distribution of dialogue moves and slots per TU.

4.3 Dialogue Behaviors

The SDN also demonstrates some interesting and unique behaviors between partners to handle unexpected situations. In particular, we investigate the
distinctive behaviors displayed by the human and the agent to handle different exceptions introduced by the Ad-Wizard. Figure 5 shows a comparison of distributions of dialogue moves and slots in EUs. The EUs are categorized by whether they handle an environmental exception, a task exception, or no exceptions introduced by the Ad-Wizard. We observe that under environmental exceptions, the agent takes more initiative to describe the situation and ask for help, with frequent use of Explain and Ask moves and use of Action and Object slots. In return, the humans initiate less Instruct moves but provide more Inform moves. Under task exceptions, humans initiate more Instruct moves with frequent use of Landmark and Status slots, in order to describe the change of plan. The agent makes confirmations with increasing use of Inform.

5 Task Definition

While many challenging tasks can be tackled using SDN, within the scope of this paper, we formulate three tasks that are critical for enabling situated dialogue for navigation. We first introduce some notations then describe our task formulation. The agent is provided with domain knowledge $K$, including a list of street names $\{s_{str_i}\}$ and (possibly incomplete) landmarks $\{l_{lm_i}\}$ on the map topology $M$. At time $t$, the interaction history (possibly empty) is represented as $H_t = \{O_t^H, (U_t^{H-1}_{1,BOT}, U_t^{H-1}_{1,BOT})\}$ which includes visual observations ($O_t^H$) and dialogue utterances from the human ($U_t^{H-1}_{1,BOT}$) and the agent ($U_t^{H-1}_{1,BOT}$). The action history, represented as $A_t$, captures the sequence of navigation actions and the dialogue moves from both the human and the agent. Given these representations, we define three tasks based on the SDN benchmark.

Dialogue Understanding for Navigation (UfN)
The UfN task challenges the agent to understand human intention (i.e., dialogue moves) from an incoming utterance. We consider each point in the SDN where the human makes an utterance as an inference point $\tau$. The task is to, at each inference point $\tau$, predict the dialogue move-slots pair $\langle d, \{s, v\}\rangle$ of the incoming utterance $u_\tau$ given the knowledge and history $\{K, M, H_\tau, A_\tau\}$.
Dialogue Moves
0.00
0.25
0.50
0.75
1.00
1.25
1.50Avg #Moves per EU
Env_HUM
Task_HUM
None_HUM
Env_BOT
Task_BOT
None_BOT
(a) Average # dialogue moves per EU.

Dialogue Slots
0.00
0.25
0.50
0.75
1.00
1.25
1.50Avg #Slots per EU
Env_HUM
Task_HUM
None_HUM
Env_BOT
Task_BOT
None_BOT
(b) Average # dialogue slots per EU.

Figure 5: The average number of dialogue moves and slots per exchange unit (EU). Different colors of the bars categorize the EUs by whether they handle an environmental exception (Env), a task exception (Task), or no exceptions (None). Darker colors indicate the move/slot is produced by the agent (BOT), and lighter colors indicate those produced by human participants (HUM).

Dialogue Response for Navigation (RfN) The RfN task challenges the agent to generate the adequate dialogue move-slot pair to drive communication. We consider each point in SDN where the Co-Wizard selected a dialogue action and/or navigation action as a decision point. The task is to, at each decision point \( \tau \), generate the dialogue move-slots pair \( \langle d, \{ s, v \} \rangle \) given the knowledge and history \( \{ K, M, H_\tau, A_\tau \} \).

Navigation from Dialogue (NfD) The NfD task challenges the agent to follow human instructions from dialogue history. The task is to, at each decision point \( \tau \) for navigation, generate the action-argument pair \( \langle p, \alpha \rangle \) for navigation given the knowledge and history \( \{ K, M, H_\tau, A_\tau \} \).

Evaluation To ensure all unexpected events and future dialogue still make sense, tasks are defined and evaluated in a teacher-forcing manner (Lamb et al., 2016; Anderson et al., 2018), e.g., the action history \( A_\tau \) presented to the model will always be the ground truth during data collection, instead of those predicted by the model at inference time. For the UfN and RfN tasks, we report the **move accuracy** and **dialogue slot F1-score** of each dialogue move and slot-value pair. In the NfD task, the argument for navigation actions is a yaw rotation angle \( \alpha \in [-180, 180] \). During evaluation, a prediction that deviates for less than 15 degrees will be considered accurate. We report the **action accuracy** with and without argument.

6 Temporally-Ordered Task-Oriented Transformer (TOTO)

Motivated by recent advances in decision-making transformers (Chen et al., 2021; Pashevich et al., 2021; Zhang and Chai, 2021), we present Temporally-Ordered Task-Oriented Transformer (TOTO), a Transformer-based baseline. TOTO is temporally-ordered as it assigns sinusoidal temporal encodings for input history instead of recurrent updates of hidden state, and is task-oriented as a unified architecture for all 3 tasks on the SDN benchmark. The text, speech, and vision inputs are each encoded using frozen pre-trained unimodal models. After encoding, the temporally encoded input is concatenated and passed through a multi-layer transformer. The output embeddings are sent to fully-connected layers to decode task outputs. The model architecture is illustrated in Figure 6, and more details are provided in Appendix C.2.

Long-Term Semantic Memory For each known landmark, we encode it with a look-up table for its location and a pre-trained BERT model (Devlin et al., 2019) for its name. While each street corresponds to a subgraph on the map topology, we first encode the map with a graph attention layer (Veličković et al., 2018), and then concatenate the pooling of each subgraph with the BERT embedding for street names. Since the knowledge \( K \) and map topology \( M \) are provided from the beginning of a session, each knowledge and street embedding is assigned with a zero temporal encoding.

Long-Term Episodic Memory We encode the complete dialogue history and action history in this module. The transcribed dialogue history, together with special speaker role tokens, is tokenized and encoded by a pre-trained BERT model. The full navigation action history, including both navigation actions and dialogue moves with their arguments and slots, is encoded with look-up tables following prior work (Pashevich et al., 2021). Each utterance and action embedding is assigned with their corresponding temporal encoding.
Short-Term Working Memory  Given the long-range nature of SDN, encoding the complete vision and trajectory history is not computationally realistic. We instead assign a fixed window size of $T = 100$ with a step $\Delta t = 4$ to sample and encode the nearest vision and trajectory history. We refer to this module as the short-term working memory encoder. Each location on the trajectory is encoded with the same look-up table in semantic memory, and each image in the visual stream is encoded with a pre-trained ResNet-50 (He et al., 2016) backbone.

Situation Awareness  Situational awareness is crucial to handle unexpected events. To this end, we attend to the current speech input with a pre-trained HuBERT (Hsu et al., 2021) encoder. In addition, we train a transformer-based object detector from $30k$ images sampled from seen splits in CARLA. The model is based on Deformable DETR (Zhu et al., 2021) and SegFormer (Xie et al., 2021), pre-trained from the supervision of bounding boxes as well as depth and semantic segmentation obtained from pseudo-sensors. The speech and object embeddings are assigned with the temporal encoding of current timestamp $t$.

7 Experiments and Results

We summarize the experiment results in Table 3. Our initial end-to-end transformer model is able to handle all tasks uniformly on both the seen and unseen splits of the test set, and outperform the majority of the unimodal baselines (See Table 5 for full results). In general, the performance is more comparable on inference tasks than decision tasks, e.g., predicting the dialogue moves from human utterances is a more approachable task than predicting the dialogue moves and navigation actions in response. We also noticed that the results on the unseen splits are uniformly better than the seen set. This can partially be explained by the fact that the unseen environment (Town 2) is significantly smaller in size and simpler in map topology. Overall, our experiment has shown that the tasks in SDN are challenging. Comparatively, the Episodic Transformer (E.T.) (Pashevich et al., 2021) baseline particularly underperforms in the inference task, i.e., UfN move prediction. The fine-tuned language model baseline can handle dialogue move predictions very well, but significantly fails on other tasks. We further provide a set of ablation studies, and discuss potential reasons why SDN is a challenge for end-to-end models. Additional results are available in Appendix C.4.

Ablation on Input Modalities  To understand how each input modality contributes to task performance, we conduct ablation studies by removing one of the input modalities. All experiments on action-level tasks (inferring or predicting dialogue moves or physical actions) are mostly influenced by the action history, which can be explained by the fact that ground truth action history is available in the teacher-forcing setup. Not surprisingly, the understanding of the incoming utterance is also largely influenced by the input of current speech. Counter-intuitively, we noticed that removing the map and knowledge encoder does not lead to decreased performance. This observation suggests that the end-to-end approach may not be reliable in tasks that require reasoning and planning over graphs, especially route planning. Overall, there is no statistically significant evidence that the full model benefits from perceptual history. We also notice that the slot-F1 scores stays relatively constant across ablation, indicating that the slot-value prediction remains challenging for TOTO.
We introduced DOROTHE, a high-fidelity simulation platform to support WoZ studies for situated communication with autonomous driving agents that can adapt to unexpected events. We defined and collected SDN, a fine-grained benchmark for continuous, dynamic, interactive navigation with sensorimotor-grounded dialogue. Our DOROTHE platform, together with our SDN benchmark, contribute a valuable resource for several lines of work in Robotics Navigation and Human-Robot Communication. We presented Temporally-Order Task Oriented Transformer (TOTO), a fully transformer-based baseline model for the SDN task. Our empirical results have shown that such long-horizon navigation tasks with rich dialogue phenomena and unexpected situations can be very challenging for end-to-end approaches. This work has shown that language-guided navigation in a highly dynamic environment (e.g., in the context of AVs) is an extremely difficult task. Our DOROTHE simulation environment, the SDN benchmark, and baseline models provide a stepping stone towards future efforts in this challenging space.

Acknowledgements

This work was supported in part by the Automotive Research Center (ARC) at the University of Michigan, NSF IIS-1949634, and NSF SES-2128623. The authors would like to thank Yichi Zhang and Jiayi Pan for their helpful discussions and the anonymous reviewers for their valuable feedback.
Limitations

While the SDN collects control streams of the vehicle, we scope our preliminary experiment on high-level robotic action prediction due to the complexity of the full problem. Specifically, our current task setup only involves the prediction of physical action and dialogue moves, and is based on several assumptions. First, there is an oracle that decides when to initiate a physical action or dialogue move. Second, we assume a ground truth low-level local planner that maps a navigation action to a sequence of control. Third, we bypass the complexity of language generation. We also ignore the adaptive and epistemic actions (speed and light changes) involved and focus on navigation. These assumptions allow us to focus on the understanding and response of situated dialogue in the navigation task and enable automatic evaluation.

In the future, we will expand the model to fully autonomous navigation settings, with sequential decision-making and hierarchical control policies. We are also interested in computational approaches for situated dialogue state tracking and management, especially involving in natural language generation and human partner modeling. With the above milestones accomplished, we will look into Sim2Real transfer and deploy our algorithm in physical autonomous driving agents.

Ethics Statement

The institution’s Institutional Review Board (IRB) considered this project exempt from ongoing review. The SDN benchmark contains human generated data (speech and demonstrations). Due to the Wizard-of-Oz nature of the study, the participants consent preceding the study and are debriefed at the end of the study. The data collection among research staff and volunteers are in line with standard ethical practice. For broader social impact, DOROTHEI aims at empowering autonomous vehicles with the ability to harness human knowledge and expertise through dialogue, and enabling natural language communication and collaboration in tackling unexpected situations. Since the dataset was developed from the simulator, the safety concerns are minimal. A complete ethical statement is available in Appendix D.

References


Harm De Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the


A Simulation Platform Setup Details

A.1 Duo-Wizard-of-Oz
The majority of the VLN benchmarks separates language from demonstration in data collection, i.e., generate the trajectory first and then annotate them with language descriptions. Our data collection, however, is inspired by the task-oriented in-vehicle dialogue corpora (Kawaguchi et al., 2004; Hansen et al., 2005; Eric et al., 2017) collected through Wizard-of-Oz (WoZ) systems (Rick, 2012). In each gameplay session, a naïve participant, who is unaware of the Wizard, will communicate with the vehicle to visit goal locations specified in a storyboard. Each task in the storyboard can potentially be changed due to unexpected alterations of the goal or environmental conditions on-the-fly. In this way, DOROTHEIE is unique in that it involves human subjects and data collection with unexpected events in naturalistic scenarios. Compared to stage-wise data collection, WoZ studies ensure synchronous and natural human-agent interaction, leading to more realistic interaction (Dahlbäck et al., 1993).

We further extend the single Wizard framework by introducing a pair of Wizards in the loop: a Collaborative Wizard (Co-Wizard) that serves the role of the original Wizard, and an Adversarial Wizard (Ad-Wizard) to control the environment and task interface to generate adversarial events on-the-fly. Without the Co-Wizard’s and participant’s awareness, the Ad-Wizard will challenge their collaboration by creating environmental changes and/or introducing task changes with appropriate context. The Co-Wizard and participant need to communicate and negotiate to arrive at an alternative plan in order to address the unexpected situations and complete the navigation tasks. The complete setup is supported by the Dialogue On the ROad To Handle Irregular Events (DOROTHEIE) platform we developed (illustrated in Figure 1).

A.2 Framework and Notations
The autonomous driving agent has a continuous control action space $\mathcal{A}$ as a triple over the normalized throttle $\phi$, steering angle $\theta$, and brake $b$, which controls the vehicle in the simulated environment. The agent is equipped with domain knowledge $K$, including a list of street names $\{\text{str}_i\}$ and (possibly incomplete) landmarks $\{\text{lm}_i\}$ on the map topology $M$. At any time $t$, the interaction history $H_t$ includes previous observation and dialogue $\{O_{t-1}^H, \{U_{t-1}^H, \text{hum}, U_{t-1}^H, \text{agent}\}\}$. The agent is, in an end-to-end manner, a system that takes $H_t$ and action history $A_t$ to produce an utterance $u_{t, \text{agent}}$ and an action $a_t \in \mathcal{A}$. Due to the long-range driving and rich interactive context, training such an end-to-end policy is unrealistic (Roh et al., 2020). Following existing systems with a modular pipeline (Skinner et al., 2016; Franke, 2017; Dosovitskiy et al., 2017), we break the problem down into perception, planning, control, and interaction, and summarize the framework in Figure 7.

Navigation Actions To navigate, the agent has a discrete and finite physical action space $\mathcal{P} = \mathcal{P}_{\text{nav}} \cup \mathcal{P}_{\text{adv}} \cup \mathcal{P}_{\text{eps}}$. Each navigation action in $\mathcal{P}_{\text{nav}} = \{\text{LaneFollow}, \text{LaneSwitch}, \text{JTurn}, \text{UTurn}, \text{Stop}, \text{Start}\}$ and its argument $\alpha \in \Theta$ is taken by the local planner to produce the continuous action $a$. The local planner defines a low-level policy $\pi \in \Pi : \mathcal{P}_{\text{nav}} \times \Theta \rightarrow \mathcal{A}$.

Adaptive Action An adaptive action $p \in \mathcal{P}_{\text{adv}} : \Pi \rightarrow \Pi$ adapts the low-level policy, e.g., by changing the target speed. In our case $\mathcal{P}_{\text{adv}} = \{\text{SpeedChange}\}$, with an increment of 5.

Sensors and Epistemic Action The sensors of the agent defines an observation function $\omega : \mathcal{S} \rightarrow \mathcal{O}$ that maps the world state $s$ to an observation $o$ (in particular, an RGB image). An epistemic action (Kirsh and Maglio, 1994) is an action taken to facilitate mental computation instead of task completion, usually by manipulating sensors, e.g., selecting active sensor types and changing camera transforms. It changes the observation func-
tion $P_{\text{eps}} : \Omega \rightarrow \Omega$. We assume for now a fixed first-person RGB camera with no noise, with only $P_{\text{eps}} = \{\text{LightChange}\}$.

**Interaction** To interact with human, the agent select a the dialogue move-slots pair $(d, \{s, v\})$, as described in Section 4.1. The language generation module generates natural language as a sequence of tokens and produces the utterance $u_{\text{agent}}$ with Google Text-to-Speech (gTTS)\(^2\).

### A.3 Simulated Environment Details

CARLA supports waypoint precision of 2cm to benchmark the continuity of robotics navigation, as well as HD Maps under the OpenDrive 1.4 Standard (Dupuis and Grezlikowski, 2006), with unique IDs for lanes, roads, and junctions to represent high-level navigation plans. To benchmark the rich environmental dynamics of outdoor navigation, we simulate multi-agent traffic with other active vehicles, bikers, and pedestrians, as well as different weather and light conditions in CARLA.

**Multi-agent Environment** Weather and light conditions can be controlled and configured on-the-fly. CARLA simulates multiple agents sharing the same environment, including vehicles and pedestrians. We use CARLA’s built-in traffic manager to simulate realistic traffic behaviour.

**Multi-sensory Perception** CARLA allows flexible sensor suites on agents, including realistic sensors (e.g., RGB, LIDAR, Radar, IMU, GNSS) and pseudo-sensors of ground truth (e.g., depth, semantic segmentation, obstacle detector). While our experiment involves only RGB images with object bounding boxes, depth, and semantic segmentation for supervision, it is possible to obtain additional sensory data for more complicated multi-modal studies.

### A.4 Navigation Task Setup

**Task Configuration** In each trial, a storyboard will specify the names and contexts of two to six landmarks to visit. Except for the final destination, the intermediate subgoals are unordered. The participant and Co-Wizard need to collaborate to guide the vehicle through all the intermediate landmarks, starting from a departure location to a destination landmark. Each session configuration is seeded from four different towns and a set of storyboard templates, with all landmark locations, street names and departure locations randomly shuffled. An example storyboard template is presented as follows:

```json
{
  "story": "Your friend Annabel is moving to a new house, and you decided to help her by doing some shopping for her. You need to get $I1 and $I2 from $P1 and $I3 from $P2, and head to Annabel’s new $P3 to help her clean the house."
}
```

```json
"subgoals": [
  {"destination": "$P1",
   "description": "Pick up $I1 and $I2 from $P1"},
  {"destination": "$P2", 
   "description": "Pick up $I3 from $P2"},
  {"destination": "$P3", 
   "description": "Arrive at Annabel’s new $P3"}
],
```

```json
"variables": 
  [["P1", "places.stores"],
   ["P2", "places.stores"],
   ["P3", "places.residential"]],
```

```json
"dependents": 
  [["I1", "P1.items"],
   ["I2", "P1.items"],
   ["I3", "P2.items"]],
```

**Knowledge Disparity** Both the Co-Wizard and the participant perceive the environment through a stream of RGB images. To replicate realistic outdoor navigation, an aerial map of the environment, with landmarks and current location, is provided to both the participant and the Co-Wizard. While both players have access to some landmarks (e.g., the location of a restaurant or grocery store), the Co-Wizard does not have access to some of the landmarks (e.g., the location of a friend’s house or a person to pick up). Such knowledge disparities motivate situated communication beyond control asymmetry and challenge the agent to understand language instructions of different granularity. For example, in the storyboard above, the Co-Wizard has no access to the location of Annabel’s new house.

```json
{
  "hidden_from_wizard": ["P3"]
}
```

### A.5 Interface Components

Figure 1 shows the conceptual overview of our interface design, including 4 major components: the Camera View, the Aerial View, the Task Interface, and the Communication Protocol. They are illustrated and described in Figure 8.

\(^2\)https://gtts.readthedocs.io/
Camera View Both the participant and the Co-Wizard have access to a first-person view of the simulated environment, similar to that of a driver.

Aerial View Both the participant and the Co-Wizard have access to a 2D aerial map of the town, with the location and heading direction of the ego-vehicle. Various annotations are included to facilitate operation, including trajectory history, street names and landmarks. The Co-Wizard is additionally shown the planned waypoints, possible trajectories at junctions, and vehicle status.

Task Interface The task interface is for human participants only. It displays the landmarks to visit as specified by the storyboard template. A subgoal is automatically fulfilled when the vehicle stops within 2 meters from the waypoint closest to the landmark, and the corresponding subgoal on the participant’s interface will turn from white to green. The participant needs to communicate with the Co-Wizard in natural language to inform the status of a subgoal. When a subgoal is added, changed, or deleted by the Ad-Wizard, the task interface will correspondingly change.

Communication Protocol Instead of typing on a keyboard, our Communication Protocol contains a dialogue interface and speech-text conversion to allow press-and-talk communication, requiring minimal distraction from operating the vehicle. This also allows the Co-Wizard to speak to the subject via speech synthesis to maintain the illusion of an automated operator to fulfill the WoZ purpose.

A.6 Co-Wizard Interface and Actions

Physical Actions We found in pilot studies that the low-level free-form controller with continuous action space is not desirable, due to the poor quality of demonstrated trajectories and high cognitive load on the Co-Wizard. On the other hand, a high-level, point-to-point maneuver controller (Dosovitskiy et al., 2017) does not support the flexibility of human command, especially in our study with heavy replanning. Motivated by prior work (Roh et al., 2020; Codevilla et al., 2018; Mueller et al., 2018), we developed a set of high-level physical actions from pilot studies for the Co-Wizard to control the vehicle. Similar to Mueller et al. (2018), we first map each action to a rule-based local trajectory planner to generate a list of waypoints that the
vehicle will drive through, then feed the waypoints to a PID controller to produce the control signals for the agent.

**Mental Actions** In a complex navigation task with multiple subgoals, belief tracking over plans, goals, task status and knowledge becomes crucial (Ma et al., 2012; Misu et al., 2014). Besides controlling the vehicle and communicating with the participant, the Co-Wizard also annotates the mental actions during and after the interaction.

- **PlanUpdate**: Indicate a change in the intended trajectory towards a destination. This is done during interaction, by noting down the navigation plan by clicking junctions on the intended trajectory from current position to the destination.
- **GoalUpdate**: Indicate the current subgoal, i.e., an intended landmark. This is done during interaction, by clicking a known landmark on the aerial view interface.
- **StatusUpdate**: Indicate a change in the belief of the task status. Since the task interface is not available to the Co-Wizard, this is done post-interaction, by annotating which participant utterances indicate a change of task status.
- **KnowledgeUpdate**: Guess the location of an unknown landmark. Since some landmarks are hidden from the Co-Wizards, they need to guess where the destination is by comprehending the participant’s descriptions. This is done during interaction, by clicking an arbitrary point on the aerial view interface.

### A.7 Ad-Wizard Interface and Actions

The Ad-Wizard is able to introduce environmental exceptions and task exceptions.

- **Environmental Exceptions**: Triggered by the change to the environment. These include direct environmental changes (Figure 10a), which challenge the vehicle’s perceptual processing and motivate participants to request for adaptations without changing the plan or goal (e.g., drive slowly in foggy weather and turn the headlights on at night). Environmental exceptions can also be introduced by creating road blocks (Figure 10b), which motivate changes of plan by failing an original navigation plan towards a landmark.

- **Task Exceptions**: Brought by changing the tasks specified in the storyboard by deleting, adding, or changing a landmark to visit (Figure 10c). The Ad-Wizard will send a message to prompt the participant in the message interface with appropriate...
context, and modify the task interface that specifies the landmarks to visit. Since the Co-Wizard does not have a task interface, the participant needs to communicate with the Co-Wizard in natural language to inform the status of a subgoal, especially when a change of current subgoal is indicated by the Ad-Wizard.

The Ad-Wizard is able to create environmental changes by modifying the weather and light conditions, and spawning more agents. To motivate changes of plan on the fly, the Ad-Wizard can create roadblocks on the paths towards a destination. Besides controlling the environment, the Ad-Wizard can also change the tasks specified in the storyboard by deleting, adding, or changing a landmark to visit. For example, in Figure 1, the Ad-Wizard attempts to change the original plan by sending a text to the human subject.

### B Dataset Details

#### B.1 Comparison of Settings

We elaborate on our comparison of settings with existing language-conditioned navigation tasks in Table 1 in the following dimensions:

- **Environment Fidelity**: The environment in which the agent operates. Depending on the fidelity, the settings are categorized into Simulated environment, Panoramic photos, and Physical environment.
- **Environment Continuity**: Whether the environment is Discrete or Continuous.
- **Turns of Communication**: Whether the communication between human and agent is Single-term or Multi-term in a task.
- **Communication Form**: Whether the form of human-agent communication is in Freeform Dialogue, Restricted Dialogue that involves turn-taking QAs, or consists of Multiple Instructions from the human only.
- **Language Granularity**: Whether the instructions are on the goal level (High) or step-by-step on the movement level (Low).
- **Control Granularity**: Whether the actions are on the discrete action level (High) or on the continuous control level (Low).
- **Language Collection**: Whether the language is collected from real Humans or generated from pre-defined Templates.
- **Demonstration Collection**: Whether the demonstrations or trajectories are collected from real Humans or generated from Planners.
- **Instruction Type**: The types of instructions initiated by a human, including possibly Replanning that requires a change of plan; Adaptation that requires adapting the manner of actions without changes of plans; Navigation that specifies a navigation action, or Manipulation that requests the agent to interact with an object.
- **Modalities**: The input modality to the agent’s sensors, including possibly Language (text), Vision (images/videos), Map, or Speech.
- **Action Space**: The output granularity of the agent’s motors, possibly Discrete or Continuous.

#### B.2 Replay and Synchronization

We use CARLA 0.9.11 with Unreal 4.24 for data collection. We apply asynchronously simulation at recording time for smooth interaction, and synchronous simulation at replaying time to retrieve sensory data at all frames without loss. For each game, we record an interaction log and a game log under a fix time step of 30 FPS with 16 substeps for physics computation, and replay the session at 10 FPS following prior work (Roh et al., 2020). The interaction log stores the timestamped activity history of the two wizards and the participant, including action history, spoken dialogue utterances, annotated mental actions, system prompts of the completion of tasks, etc. The game log stores the world state at each timestamp to reproduce a game, including locations, orientations, bounding boxes, velocity, and physical control signals of vehicles and states of traffic lights, etc. By attaching sensors to the ego-vehicle in replay, we are able to log RGB perception streams together with the ground truth pseudo-sensors (depth and semantic segmentation).

#### B.3 Dialogue Annotation

**Annotating Transcripts** The trimmed audio clips are first sent to Google Speech Recognition for raw text. We then listen to each trimmed clip and type the ground truth transcripts. Based on the ground truth transcripts, we further annotate each dialogue session using four levels of linguistic units, described as follows.

**Annotating Transactions** *Transaction Units (TUs)* are sub-dialogues that starts when a task is...
Annotating Exchanges  Exchange Units (EUs) are sequences of dialogue moves towards common ground. They start with an initiating utterance that has a purpose (e.g., a question) and ends when the expectations are fulfilled or abandoned (e.g., an answer). We observed in the corpus that some exchange units overlap because the participant and the Co-Wizard spoke up at the same time. This is particularly common when the Co-Wizard initiates a conversation asking for instructions and meanwhile the participant is giving the command. The annotators are tasked to match each utterance to an exchange unit, represented by its initiating utterance.

Annotating Dialogue Moves  Dialogue Moves are sub-categories of dialogue acts that drive conversation and update domain-specific information state within an exchange. We follow the coding scheme of Carletta et al. (1997) to represent dialogue moves as a decision tree, with a slight modification to adjust to our domain ontology, as presented in Figure 2. The 14 dialogue moves, together with Irrelevant, specify the space of conversational action in the human-vehicle dialogue. The annotators are tasked to first split each utterance into text spans that contain only one dialogue move, and then assign the move to the span following the decision tree.

Annotating Dialogue Slots  Dialogue Slots are parameters that further determine the semantics of dialogue moves. We consider 5 slot labels: Action, Street, Landmark, Status, Object. For each slot label, the slot value belongs to a finite set of possible values defined by the domain ontology, e.g., the Action is specified by the physical action space, and the Object is specified by CARLA’s built-in visual semantics.

• 10 values for Action: Queried, Unknown, LaneFollow, LaneSwitch, JTurn, UTurn, Stop, Start, SpeedChange, LightChange.
• 17 values for Street: Queried, Unknown, Baits, Beal, Bishop, Bonisteel, Broadway, Division, Draper, Duffield, Fuller, Hayward, Hubbard, Murfin, Plymouth, Upland, Highway.
• 12 values for Landmark: Queried, Unknown, BurgerKing, Coco, Ikea, KFC, Panera, Qdoba, SevenEleven, Shell, House, Person.
• 6 values for Status: Queried, Unknown, Ongoing, Complete, Abandoned, Pending.
• 24 values for Object: Queried, Unlabeled, Building, Fence, Pedestrian, Pole, RoadLine, Road, SideWalk, Vegetation, Vehicles, Wall, TrafficSign, Sky, Ground, Bridge, RailTrack, GuardRail, TrafficLight, Static, Dynamic, Water, Terrain, Other.

Kappa Analysis  We report the average Cohen’s kappa (κ) (Cohen, 1960) for each pair of annotators sharing common judgements in 20 common sessions with 1045 utterances. We obtain κ = 0.77 ± 0.02 for dialogue move annotation, and κ = 0.85 ± 0.02 for dialogue slot annotation.

B.4 Additional Dialogue Samples

The human and the agent communicate about objects in the scene, and the human frequently checks the perceptual capability of the agent.

[Human] watch out for the biker
[Human] let's be careful
[Bot] okay I’ll keep it in mind
[Human] yep... now slowly
[Bot] okay
[Human] are you able to see the biker ahead
[Bot] yes I can see the biker ahead

When a new task is introduced, the human and the agent negotiate towards a mutually agreed plan before the agent takes action.

[Human] do you see the people on the other side of the road
[Human] that’s my friend
[Human] let’s go pick him up
[Bot] okay
[Bot] I think I need to be coming from Broadway Street to go on the correct side of the road
[Bot] would you like for me to do that
[Human] okay then make a right turn
[Human] and then make a right turn at Murfin
[Bot] okay I’ll do that
[Human] yes please let’s go
When the human gets confused by the agent’s action, they would ask the agent to describe and explain its plan. This requires the agent to reason about the route planning task, and generate language description of it.

Human: what are you doing
Bot: I'll continue going straight and then at the roundabout
Human: why don't you make a right turn here
Bot: I won't be able to go on the correct side of the road to pick up your friend
Human: I actually think I have to come from south on Bonisteel Avenue for that
Bot: would you like for me to do that instead
Human: no just do what you had in mind before

Episodic memory is also crucial to complete the tasks in SDN. The agent needs to keep track of the dialogue and visual history in order to resume a previously abandoned task.

Human: my aunt's house is going to be on the right side
Bot: okay could you tell me when to stop on Beal Street

<Ad-Wizard> You forgot your bag in the KFC>

Human: let's go back to KFC
Bot: okay I'll go back to KFC
Human: I think we're at the KFC now
Bot: okay great now we're going to go back to my aunt's house
Human: okay
Bot: my aunt's house is at the Baits and Beal’s intersection
Human: so let's take a right turn at Baits
Bot: okay I'll do that
Human: we went to my aunt's house before
Bot: do you remember where it is
Human: yes I do
Bot: what's the location of my aunt's house
Human: you're aunt's house is at the intersection of Baits Avenue and Beal Street on the right side of the Beal Street once I turn right onto the Beal intersection of Baits Avenue and Beal South on Bonisteel Avenue for that
Bot: that's correct that's where we're going right now
Human: okay I'll take you there
Bot: and once we get there we're just going to park on the side of the road
Human: okay

C Experiment Details

C.1 Computational Resources

The recording of all sessions is done on a machine with an NVIDIA A4000 GPU. The TOTO and baseline models are trained on one NVIDIA A40 GPU. For each experiment we train at least 10 epochs with early stop of 5 epochs, which take two hours in average.

C.2 Baseline Details

The complete model is illustrated in Figure 11.

Decision Transformer The encoding transformer has two layers each with 11 heads and dropout rate with 0.1. The length of input feature dimension is 770 and we use the sin-cos function along the whole game as temporal embedding of history. We used frozen BERT and HuBERT as our encoders for text and speech. To process the image and extract features, we train a model using Segformer and Deformable-DETR. We first trained a Segformer model to predict depth and semantic segmentation using the RGB images generated by the simulator. Adding these two new channels into the RGB image, we augment the incoming image into 5 channels. We then develop a Deformable-DETR model with 300 object queries using the 5-channel images as input, and obtain an embedding with a size of $300 \times 256$ for each image.

C.3 Hyper-Parameter Decisions

We include major hyper-parameter decisions for reproducibility purposes. Please refer to supplementary code for more details.

Depth and Semantic Prediction (Segformer) With the same backbone as Segformer, another head is added to predict the depth along with the semantic segmentation.

- learning rate: $6 \times 10^{-5}$
- weight decay: 0.01
- semantic class number: 23
- depth class number: 256
- optimizer: AdamW

Object Detection (Deformable-DETR) The original Deformable-DETR receives an RGB image input, while we have 5-channel (RGB, Depth, Semantics) augmented images. We first use a fully connected layer to encode the input and then use the original Deformable-DETR model. Two heads are used to predict the 2D and 3D bounding box for each query during training.

- learning rate: $2 \times 10^{-4}$
- backbone learning rate: $2 \times 10^{-5}$
- weight decay: $1 \times 10^{-4}$
- optimizer: AdamW
Learning Parameters

- learning rate: $1 \times 10^{-5}$
- weight decay: $1 \times 10^{-4}$
- optimizer: AdamW

Loss coefficients

- location belief MSE loss: $1 \times 10^{-3}$
- UfN dialogue move type cross entropy loss: $1$
- UfN dialogue slot type BCE loss: $2 \times 10^{-2}$
- UfN dialogue slot value BCE loss: $2 \times 10^{-3}$
- RfN dialogue move type cross entropy loss: $1$
- RfN dialogue slot type BCE loss: $1 \times 10^{-2}$
- RfN dialogue slot value BCE loss: $2 \times 10^{-3}$
- NfD Physical action type cross entropy loss: $1$
- NfD Physical action arg MSE loss: $1 \times 10^{-2}$

C.4 Addendum to Results

Vision Model Performance  The performance of the vision models on semantic segmentation and object detection are provided in Table 4. While the performance itself is competent, the decision transformer fails to benefit from the vision representations. This observation indicates that reasoning over cross-modal information may be a bottleneck for current end-to-end baselines.

<table>
<thead>
<tr>
<th>Semantic Segmentation</th>
<th>Object Detection</th>
</tr>
</thead>
<tbody>
<tr>
<td>mIoU</td>
<td>AP(_{b}d=0.50)</td>
</tr>
<tr>
<td>65.60%</td>
<td>71.40%</td>
</tr>
<tr>
<td>mAcc</td>
<td>AP(_{b}d=0.50;0.95)</td>
</tr>
<tr>
<td>71.60%</td>
<td>41.30%</td>
</tr>
</tbody>
</table>

Table 4: The performance of the vision models on semantic segmentation and object detection.

D Ethical Considerations

D.1 Consent Statement

You are invited to participate in a research study that intends to develop approaches to support language communication between humans and autonomous vehicles (AV). If you agree to be part of the research study, you will be asked to interact with a simulated AV in a virtual world to accomplish a set of tasks. Imagine you need to send your car out to do some errands. While the car is out, there may be some unexpected situations happening (e.g., a fallen tree blocks the road, you need to add a stop, etc.) You will need to communicate with your car in natural language to help the car deal with these exceptions and achieve the tasks. The study will last approximately an hour. The interaction between you and the car (i.e., speech/chat) and screen activities (i.e., the movement of the car in the virtual environment and its surroundings) will be recorded in a datafile. The data collected in this study will be analyzed and used for research purposes. No personally identifiable information besides the audio recording will be stored in the datafile.

D.2 Debriefing Statement

Earlier in the consent form, we informed you that you will be asked to interact with a simulated AV in a virtual world to accomplish a set of tasks. In actuality, the vehicle is not controlled by an algorithm, but by our research staff. The exceptions you encountered during the study were generated by our research staff on the fly. Unfortunately, due to the nature of this Wizard-of-Oz study, we could not provide you with all of these details prior to your participation. This ensures that your reactions in this study were spontaneous and not influenced.
<table>
<thead>
<tr>
<th>Model</th>
<th>UN (Seen)</th>
<th></th>
<th>RFN (Seen)</th>
<th></th>
<th>NFD (Seen)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTO</td>
<td>40.9±3.9</td>
<td>36.9±0.0</td>
<td>29.2±0.7</td>
<td>55.7±0.2</td>
<td>41.2±2.5</td>
<td>36.0±3.4</td>
</tr>
<tr>
<td>TOTO (+ Belief Tracking)</td>
<td>39.5±2.2</td>
<td>37.0±0.1</td>
<td>28.8±0.9</td>
<td>55.7±0.2</td>
<td>40.7±3.6</td>
<td>34.0±4.7</td>
</tr>
<tr>
<td>TOTO (+ Fine-tuned BERT)</td>
<td>38.4±2.7</td>
<td>36.9±0.0</td>
<td>27.8±0.7</td>
<td>55.7±0.0</td>
<td>43.6±1.6</td>
<td>30.9±4.1</td>
</tr>
<tr>
<td>TOTO (- Action History)</td>
<td>30.5±1.5</td>
<td>36.9±0.0</td>
<td>23.5±1.7</td>
<td>55.7±0.0</td>
<td>27.6±2.8</td>
<td>24.6±4.0</td>
</tr>
<tr>
<td>TOTO (- GT Transcript)</td>
<td>39.8±1.9</td>
<td>36.9±0.1</td>
<td>29.2±0.8</td>
<td>55.6±0.1</td>
<td>40.4±3.4</td>
<td>31.6±4.3</td>
</tr>
<tr>
<td>TOTO (- Object Detection)</td>
<td>42.5±2.8</td>
<td>37.0±0.2</td>
<td>30.4±0.7</td>
<td>55.8±0.1</td>
<td>39.2±3.5</td>
<td>34.4±5.8</td>
</tr>
<tr>
<td>TOTO (- Vision History)</td>
<td>41.9±1.3</td>
<td>37.0±0.2</td>
<td>29.1±0.5</td>
<td>55.8±0.2</td>
<td>42.0±3.1</td>
<td>36.1±4.0</td>
</tr>
<tr>
<td>TOTO (- Current Speech)</td>
<td>35.1±2.7</td>
<td>36.7±0.5</td>
<td>29.9±0.9</td>
<td>55.9±0.2</td>
<td>39.7±1.9</td>
<td>33.7±3.0</td>
</tr>
<tr>
<td>TOTO (- Map Knowledge)</td>
<td>42.6±1.2</td>
<td>36.9±0.0</td>
<td>29.3±0.9</td>
<td>55.8±0.2</td>
<td>44.6±3.3</td>
<td>39.1±3.3</td>
</tr>
<tr>
<td>Fine-tuned BERT Only</td>
<td>66.8±2.0</td>
<td>24.9±5.5</td>
<td>52.7±1.0</td>
<td>46.0±2.5</td>
<td>32.4±1.2</td>
<td>16.2±2.7</td>
</tr>
<tr>
<td>BERT Only</td>
<td>52.1±3.2</td>
<td>39.9±1.3</td>
<td>52.3±1.0</td>
<td>56.1±0.3</td>
<td>30.1±1.8</td>
<td>25.6±2.8</td>
</tr>
<tr>
<td>HuBERT Only</td>
<td>35.7±3.4</td>
<td>36.9±0.0</td>
<td>24.5±0.1</td>
<td>55.6±0.1</td>
<td>31.2±0.9</td>
<td>31.1±0.9</td>
</tr>
<tr>
<td>Deformable-DETR Only</td>
<td>31.5±0.0</td>
<td>36.9±0.0</td>
<td>24.4±0.0</td>
<td>55.7±0.0</td>
<td>31.5±0.6</td>
<td>31.5±0.6</td>
</tr>
<tr>
<td>Map Encoder Only</td>
<td>25.3±2.5</td>
<td>36.9±0.0</td>
<td>22.5±1.6</td>
<td>55.7±0.0</td>
<td>29.7±0.6</td>
<td>28.4±1.8</td>
</tr>
<tr>
<td>E.T.</td>
<td>36.6±3.6</td>
<td>37.0±0.2</td>
<td>29.4±1.2</td>
<td>55.9±0.2</td>
<td>40.0±2.8</td>
<td>32.2±4.0</td>
</tr>
<tr>
<td>E.T. (+ Fine-tuned BERT)</td>
<td>33.6±1.1</td>
<td>36.8±1.0</td>
<td>26.5±0.9</td>
<td>55.7±0.0</td>
<td>38.0±1.1</td>
<td>27.6±6.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model</th>
<th>UN (Unseen)</th>
<th></th>
<th>RFN (Unseen)</th>
<th></th>
<th>NFD (Unseen)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTO</td>
<td>49.2±3.0</td>
<td>26.2±0.0</td>
<td>31.0±1.7</td>
<td>54.0±0.7</td>
<td>45.8±3.8</td>
<td>41.1±2.8</td>
</tr>
<tr>
<td>TOTO (+ Belief Tracking)</td>
<td>47.1±3.5</td>
<td>26.2±0.0</td>
<td>29.0±2.0</td>
<td>53.7±0.7</td>
<td>47.6±4.5</td>
<td>38.8±4.1</td>
</tr>
<tr>
<td>TOTO (+ Fine-tuned BERT)</td>
<td>49.6±0.9</td>
<td>26.2±0.1</td>
<td>34.0±2.1</td>
<td>54.8±0.0</td>
<td>48.5±4.6</td>
<td>36.5±5.5</td>
</tr>
<tr>
<td>TOTO (- Action History)</td>
<td>35.5±3.2</td>
<td>26.1±0.1</td>
<td>28.2±3.9</td>
<td>54.8±0.0</td>
<td>36.8±0.8</td>
<td>36.0±1.7</td>
</tr>
<tr>
<td>TOTO (- GT Transcript)</td>
<td>46.7±2.4</td>
<td>26.2±0.0</td>
<td>31.6±2.6</td>
<td>54.2±0.8</td>
<td>46.2±5.9</td>
<td>37.6±6.9</td>
</tr>
<tr>
<td>TOTO (- Object Detection)</td>
<td>50.0±1.8</td>
<td>26.2±0.1</td>
<td>32.7±2.2</td>
<td>53.8±1.2</td>
<td>45.7±5.2</td>
<td>40.3±5.4</td>
</tr>
<tr>
<td>TOTO (- Vision History)</td>
<td>48.7±2.3</td>
<td>26.2±0.1</td>
<td>31.5±2.9</td>
<td>54.3±0.7</td>
<td>45.9±4.2</td>
<td>42.3±3.5</td>
</tr>
<tr>
<td>TOTO (- Current Speech)</td>
<td>42.8±2.5</td>
<td>25.8±0.3</td>
<td>33.8±1.4</td>
<td>55.1±0.4</td>
<td>46.5±4.9</td>
<td>39.4±5.2</td>
</tr>
<tr>
<td>TOTO (- Map Knowledge)</td>
<td>48.2±1.0</td>
<td>26.2±0.1</td>
<td>31.9±1.2</td>
<td>54.9±0.8</td>
<td>51.7±3.4</td>
<td>46.0±4.0</td>
</tr>
<tr>
<td>Fine-tuned BERT Only</td>
<td>67.2±1.5</td>
<td>16.2±3.5</td>
<td>57.0±0.9</td>
<td>46.9±2.2</td>
<td>37.1±1.5</td>
<td>19.8±3.6</td>
</tr>
<tr>
<td>BERT Only</td>
<td>57.3±2.1</td>
<td>31.7±2.1</td>
<td>57.4±1.3</td>
<td>55.9±0.6</td>
<td>35.3±3.6</td>
<td>30.1±4.2</td>
</tr>
<tr>
<td>HuBERT Only</td>
<td>40.9±2.3</td>
<td>26.2±0.0</td>
<td>30.2±0.2</td>
<td>54.7±0.1</td>
<td>36.7±0.9</td>
<td>36.7±0.2</td>
</tr>
<tr>
<td>Deformable-DETR Only</td>
<td>37.6±0.0</td>
<td>26.2±0.0</td>
<td>30.7±0.0</td>
<td>54.8±0.0</td>
<td>36.9±0.1</td>
<td>36.9±0.1</td>
</tr>
<tr>
<td>Map Encoder Only</td>
<td>30.1±0.9</td>
<td>26.2±0.0</td>
<td>23.1±1.0</td>
<td>54.8±0.0</td>
<td>37.0±0.1</td>
<td>37.0±0.1</td>
</tr>
<tr>
<td>E.T.</td>
<td>45.1±3.8</td>
<td>26.1±0.1</td>
<td>33.4±2.2</td>
<td>54.7±2.8</td>
<td>46.6±3.3</td>
<td>37.0±5.9</td>
</tr>
<tr>
<td>E.T. (+ Fine-tuned BERT)</td>
<td>41.8±0.6</td>
<td>26.2±0.0</td>
<td>34.1±1.1</td>
<td>54.8±0.0</td>
<td>44.5±3.1</td>
<td>32.9±7.3</td>
</tr>
</tbody>
</table>

Table 5: Complete experiment results of TOTO and the baselines.
by prior knowledge that you are interacting with another human player. We regret the deception in the study, but we hope you understand the reason for it. Knowing all of the study information, you may withdraw your data without penalty or loss of benefits to which you are otherwise entitled.

D.3 Complete Ethical Statement

The SDN benchmark contains human generated data (speech and demonstrations). The institution’s Institutional Review Board (IRB) considered this project exempt from ongoing review. Preceding each study, a staff will describe the capability of the vehicle to the participant. The participant would sign a consent form in which they were asked to communicate with a fully-functional autonomous vehicle in a simulated environment and navigate to specified landmarks. After the last session, the staff representative would debrief the participant by introducing the Wizard-of-Oz nature of the study. Participants would sign another debriefing form to indicate their voluntary agreement to participate in this study. Since some acquaintance with the DOROTHIE interface is required to play the wizard role, we trained five research staff on both the Co-Wizard and Ad-Wizard interface and they alternated in the human studies. The data collection among research staff and volunteers are in line with standard ethical practice.

D.4 Broader Impact

For broader social impact, DOROTHIE aims at empowering autonomous vehicles with the ability to harness human knowledge and expertise through dialogue, and enabling natural language communication and collaboration in tackling unexpected situations. Since the dataset was developed from the simulator, the safety concerns are minimal.