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

Ziqiao Ma[1], **Ben VanDerPloeg**[*1], **Cristian-Paul Bara**[*†1], **Yidong Huang**[*1],
Eui-In Kim[1], **Felix Gervits**[2], **Matthew Marge**[2], **Joyce Chai**[1]
[1]University of Michigan [2]U.S. Army Research Laboratory
{marstin,bensvdp,cpbara,owenhji,euiink,chaijy}@umich.edu
{felix.gervits,matthew.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].

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-

[*]Equal contribution.

[†]Work done prior to joining Amazon Alexa AI

[1]The DOROTHIE platform, SDN benchmark, and code for the baseline model are available at https://github.com/sled-group/DOROTHIE

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4800–4822
December 7-11, 2022 ©2022 Association for Computational Linguistics

Name	Domain	Environment		Communication		Granularity		Data Collection			Instruction Type				Action Space
		Fidelity	Continuity	Turn	Form	Language	Control	Lang.	Demo.	Modal.	Replan.	Adp.	Nav.	Man.	Continuity
SDN (Ours)		Sim	C	M	Freeform	H & L	H & L	H	H	LVMS	✔	✔	✔	-	D & C
CDNLI (Roh et al., 2020)	Outdoors	Sim	C	M	Multi Inst	L	H & L	H+T	P	LVM	-	✔	✔	-	D & C
LCSD (Sriram et al., 2019)		Sim	C	S	Multi Inst	L	H	H	P	LVM	-	-	✔	-	D
TtW (De Vries et al., 2018)		Pano	D	M	Freeform	H & L	H	H	H	LVM	-	-	✔	-	D
Talk2Nav (Vasudevan et al., 2021)		Pano	D	S	Multi Inst	L	H	H	P	LVM	-	-	✔	-	D
TouchDown (Chen et al., 2019)		Pano	D	S	Multi Inst	L	H	H	P	LVM	-	-	✔	-	D
Street Nav (Hermann et al., 2020)		Pano	D	M	Multi Inst	L	H	T	P	LVM	-	-	✔	-	D
Map2Seq (Schumann and Riezler, 2021)		Pano	D	S	Multi Inst	L	H	H	P	LM	-	-	✔	-	D
RUN (Paz-Argaman and Tsarfaty, 2019)		Pano	D	S	Multi Inst	L	H	H	H	LM	-	-	✔	-	D
TEACh (Padmakumar et al., 2022)	Indoors	Sim	C	M	Freeform	H & L	H	H	H	LV	-	✔	✔	✔	D
DialFRED (Gao et al., 2022)		Sim	C	M	Restricted	H & L	H	H+T	P	LV	-	✔	✔	✔	D
ALFRED (Shridhar et al., 2020)		Sim	C	S	Multi Inst	H & L	H	H	P	LV	-	✔	✔	✔	D
HANNA (Nguyen and Daumé III, 2019)		Pano	D	M	Multi Inst	H & L	H	H	P	LV	-	✔	✔	-	D
RobotSlang (Banerjee et al., 2020)		Phy	C	M	Freeform	H & L	H	H	P	LV	-	-	✔	-	D
TtT and WtW (Ilyevsky et al., 2021)		Phy	C	S	Restricted	H & L	H	H	P	LM	-	-	✔	-	D
Robo-VLN (Irshad et al., 2021)		Pano	C	S	Multi Inst	L	H & L	H	P	LV	-	-	✔	-	C
VLN-CE (Krantz et al., 2020)		Pano	C	S	Multi Inst	L	H	H	P	LV	-	-	✔	-	D
CVDN (Thomason et al., 2020)		Pano	D	M	Restricted	L	H	H	H	LV	-	-	✔	-	D
R2R (Anderson et al., 2018)		Pano	D	S	Multi Inst	L	H	H	P	LV	-	-	✔	-	D

Table 1: Comparison of language-conditioned task completion settings in terms of **Environment Fidelity** (Simulated, Panoramic, Physical), **Environment Continuity** (Discrete, Continuous), **Turns of Communication** (Single, Multiple), **Communication Form** (Freeform Dialogue, Restricted Dialogue, Multiple Instructions), **Language Granularity** (High: Goal, Low: Step/Movement), **Control Granularity** (High: Action, Low: Control), **Language Collection** (Human, Templated), **Demonstration Collection** (Human, Planner), **Modalities** (Language, Vision, Map, Speech), **Instruction Type** (Replanning, Adaptation, Navigation, Manipulation), **Action Space** (Discrete, Continuous).

nals) as well as fine-grained annotation of dialogue phenomena at multiple levels. SDN challenges autonomous driving agents to navigate in continuous and dynamic environments, engage in situated communication with humans, and handle unexpected events on the fly. As an initial step, we developed the Temporally-Ordered Task Oriented Transformer (TOTO), a transformer-based baseline model for three tasks: (1) predicting dialogue moves from human utterances; (2) generating dialogue moves in response to humans; and (3) generating navigation actions towards the goal. We present our empirical results and discuss key challenges and opportunities.

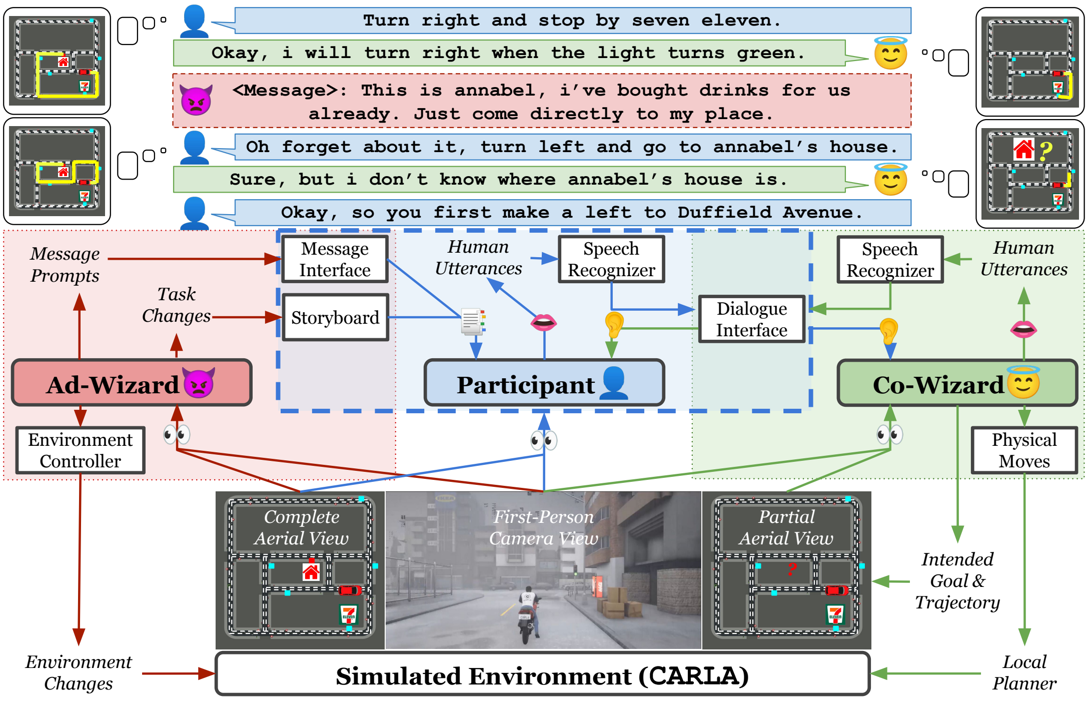
To the best of our knowledge, this is the first effort on language communication under unexpected situations in autonomous vehicles. Our contributions are the following: (1) a novel, high-fidelity simulation platform, DOROTHIE, that can be used to create unexpected situations on the fly during human-agent communication, (2) a fine-grained benchmark, SDN, for continuous, dynamic, interactive navigation with sensorimotor-grounded dialogue, and (3) a transformer-based model for action prediction and decision-making which serves as a baseline for future development.

2 Related Work

Our work is mostly related to *language-conditioned navigation* tasks (Anderson et al., 1991; MacMahon et al., 2006; Paz-Argaman 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 tasked 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.



Turn right and stop by seven eleven.

Okay, i will turn right when the light turns green.

<Message>: This is annabel, i've bought drinks for us already. Just come directly to my place.

Oh forget about it, turn left and go to annabel's house.

Sure, but i don't know where annabel's house is.

Okay, so you first make a left to Duffield Avenue.

Figure 1: An overview of the DOROTHIE design. We extend the traditional Wizard-of-Oz framework by introducing a pair of Wizards: **Co-Wizard** and **Ad-Wizard**. A Human **Participant** is given a storyboard and is instructed to communicate with an autonomous vehicle to complete a set of tasks. The Co-Wizard controls the agent's behaviors and communicates with the human. The Ad-Wizard creates unexpected situations on the fly. The human and the Co-Wizard need to collaborate with each other to resolve these unexpected situations.

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.

Physical Actions	**Args**	**Descriptions**
`LaneFollow`	-	Default behaviour, follow the current lane.
`LaneSwitch`	Angle (`Rotation`)	Switch to a neighboring lane.
`JTurn`	Angle (`Rotation`)	Turn to a connecting road at a junction.
`UTurn`	-	Make a U-turn to the opposite direction.
`Stop`	-	Brake the vehicle manually.
`Start`	-	Start the vehicle manually.
`SpeedChange`	Speed (±5)	Change the desired cruise speed by 5 km/h.
`LightChange`	Light State (`On/Off`)	Change the front light state.
Mental Actions	**Args**	**Descriptions**
`PlanUpdate`	List[`Junction ID`]	Indicate intended trajectory towards a destination.
`GoalUpdate`	List[`Landmark`]	Indicate current goal as an intended landmark.
`StatusUpdate`	Tuple[`Landmark,Status`]	Indicate a change in task status.
`KnowledgeUpdate`	`x,y`	Guess the location of an unknown landmark.
`Other`	-	Other belief state updates.

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 `DOROTHIE` 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 `DOROTHIE`, 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,

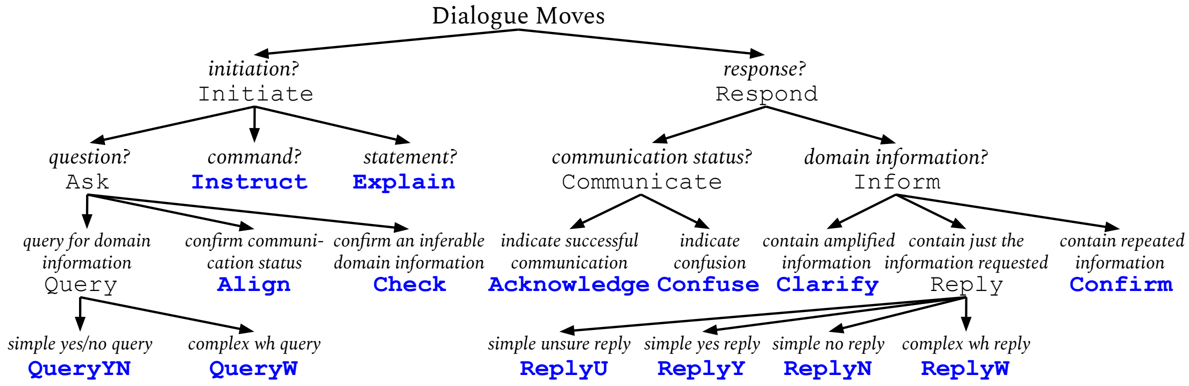


Figure 2: The coding scheme of dialogue moves as a decision tree. The leaf nodes of the decision tree specify the set of dialogue moves we used for annotation.

a continuous trajectory, and control signals. The benchmark also includes dialogue structure annotation, which we analyzed for dialogue behaviors.

4.1 Dialogue Structure Annotation

Following prior work in human-robot dialogue (Marge et al., 2017; Traum et al., 2018; Marge et al., 2020) and dialogue discourse processing (Sinclair et al., 1975; Grosz and Sidner, 1986; Clark, 1996), we annotate each dialogue session using four levels of linguistic units:

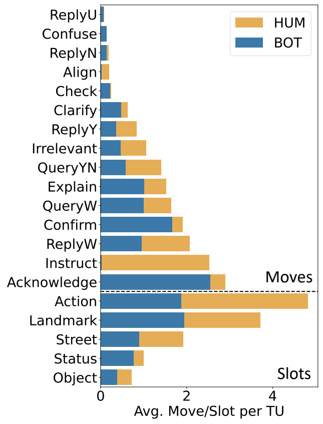
- **Transaction Units (TUs)**: Sub-dialogues that start when a task is initiated and end when it is completed, interrupted, or abandoned.
- **Exchange Units (EUs)** Sequences of dialogue moves towards common ground. These start with an initiating utterance that has a purpose (*e.g.,* a question) and end when the expectations are fulfilled or abandoned (*e.g.,* an answer).
- **Dialogue Moves** Sub-categories of dialogue acts that drive conversation and update domain-specific information state within an exchange.
- **Dialogue Slots** Parameters that further determine the semantics of dialogue moves, including Action, Street, Landmark, Status, Object.

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, as presented in Figure 2. The 14 dialogue moves, together with Irrelevant, specify the space of conversational action in the human-vehicle dialogue. We present an example dialogue with annotations in Figure 4, with more samples available in Appendix B.4.

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.

Metric	Value
Control Stream	18.7 h
Trimmed Audio	2.9 h
# Utterances	8415
# Words	50398
Vocabulary	1373
# Transactions	578
# Exchanges	4089
# Dialogue Moves	11623
# Slot Values	8618
# Physical Actions	9448
Fold (Split)	**# Sessions**
Train	123
Val (Seen)	14
Val (Unseen)	6
Test (Seen)	25
Test (Unseen)	15



(a) Dataset Statistics and split information.

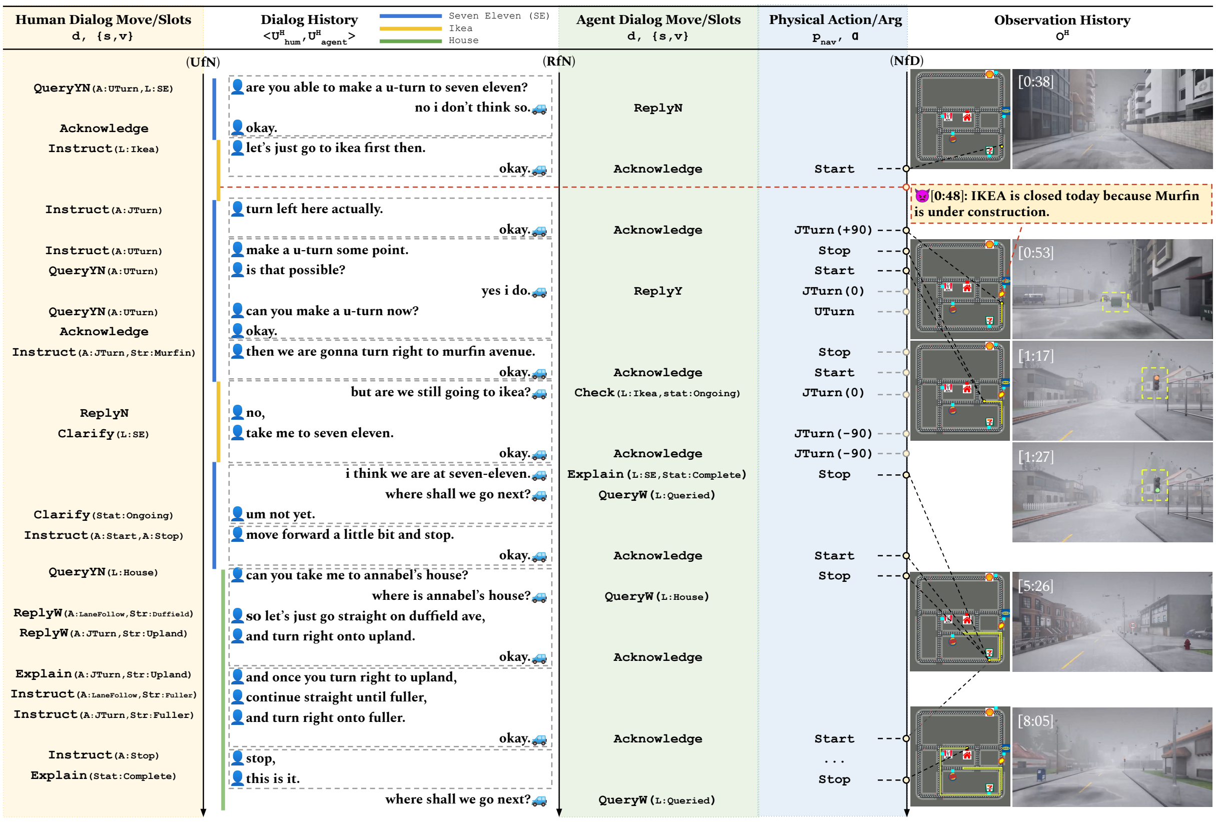
(b) The distribution of dialogue moves and slots per TU.

Figure 3: Dataset description.

Figure 3b shows the frequencies of dialogue moves and slots taken by the human and the agent respectively. Not surprisingly, due to the nature of the joint tasks, the human mostly instructs and the agent constantly provides acknowledgement and confirmation. Both the human and the agent ask questions and give answers. The agent appears to provide more explanation about its own behaviors and decisions.

4.3 Dialogue Behaviors

The SDN also demonstrates some interesting and unique behaviors between partners to handle unexpected situations. In particular, we investigate the



Human Dialog Move/Slots $d, \{s,v\}$	Dialog History $\langle U^H_{hum}, U^H_{agent}\rangle$	Agent Dialog Move/Slots $d, \{s,v\}$	Physical Action/Arg p_{nav}, a
QueryYN(A:UTurn,L:SE)	👤are you able to make a u-turn to seven eleven?		
	no i don't think so.🚙	ReplyN	
Acknowledge	👤okay.		
Instruct(L:Ikea)	👤let's just go to ikea first then.		
	okay.🚙	Acknowledge	Start
	😈[0:48]: IKEA is closed today because Murfin is under construction.		
Instruct(A:JTurn)	👤turn left here actually.		
	okay.🚙	Acknowledge	JTurn(+90)
Instruct(A:UTurn)	👤make a u-turn some point.		Stop
QueryYN(A:UTurn)	👤is that possible?		Start
	yes i do.🚙	ReplyY	JTurn(0)
QueryYN(A:UTurn)	👤can you make a u-turn now?		UTurn
Acknowledge	👤okay.		
Instruct(A:JTurn,Str:Murfin)	👤then we are gonna turn right to murfin avenue.		Stop
	okay.🚙	Acknowledge	Start
	but are we still going to ikea?🚙	Check(L:Ikea,stat:Ongoing)	JTurn(0)
ReplyN	👤no,		
Clarify(L:SE)	👤take me to seven eleven.		JTurn(-90)
	okay.🚙	Acknowledge	JTurn(-90)
	i think we are at seven-eleven.🚙	Explain(L:SE,Stat:Complete)	Stop
	where shall we go next?🚙	QueryW(L:Queried)	
Clarify(Stat:Ongoing)	👤um not yet.		
Instruct(A:Start,A:Stop)	👤move forward a little bit and stop.		
	okay.🚙	Acknowledge	Start
QueryYN(L:House)	👤can you take me to annabel's house?		Stop
	where is annabel's house?🚙	QueryW(L:House)	
ReplyW(A:LaneFollow,Str:Duffield)	👤so let's just go straight on duffield ave,		
ReplyW(A:JTurn,Str:Upland)	👤and turn right onto upland.		
	okay.🚙	Acknowledge	
Explain(A:JTurn,Str:Upland)	👤and once you turn right to upland,		
Instruct(A:LaneFollow,Str:Fuller)	👤continue straight until fuller,		
Instruct(A:JTurn,Str:Fuller)	👤and turn right onto fuller.		
	okay.🚙	Acknowledge	Start
Instruct(A:Stop)	👤stop,		...
Explain(Stat:Complete)	👤this is it.		Stop
	where shall we go next?🚙	QueryW(L:Queried)	

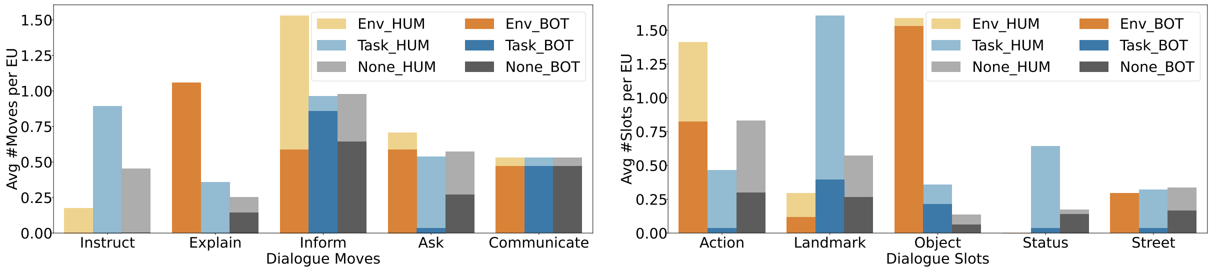
Figure 4: A simple exemplar session in (SDN) with annotations. Each color bar represents a transaction unit and each box represents an exchange unit. The tasks challenge the agent to understand input dialogue move and imitate Co-Wizard's decision on the next navigation action and dialogue move to take.

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 $\{\mathtt{str}_i\}$ and (possibly incomplete) landmarks $\{\mathtt{lm}_i\}$ on the map topology M. At time t, the *interaction history* (possibly empty) is represented as $H_t = \{O^H_{t-1}, \langle U^H_{t-1,\text{HUM}}, U^H_{t-1,\text{BOT}}\rangle\}$ which includes visual observations (O^H_{t-1}) and dialogue utterances from the human ($U^H_{t-1,\text{HUM}}$) and the agent ($U^H_{t-1,\text{BOT}}$). 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*** τ. The task is to, at each inference point τ, predict the dialogue move-slots pair $\langle d, \{s, v\}\rangle$ of the incoming utterance u_τ given the knowledge and history $\{K, M, H_\tau, A_\tau\}$.



(a) Average # dialogue moves per EU.

(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 τ, 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 τ 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_τ 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.

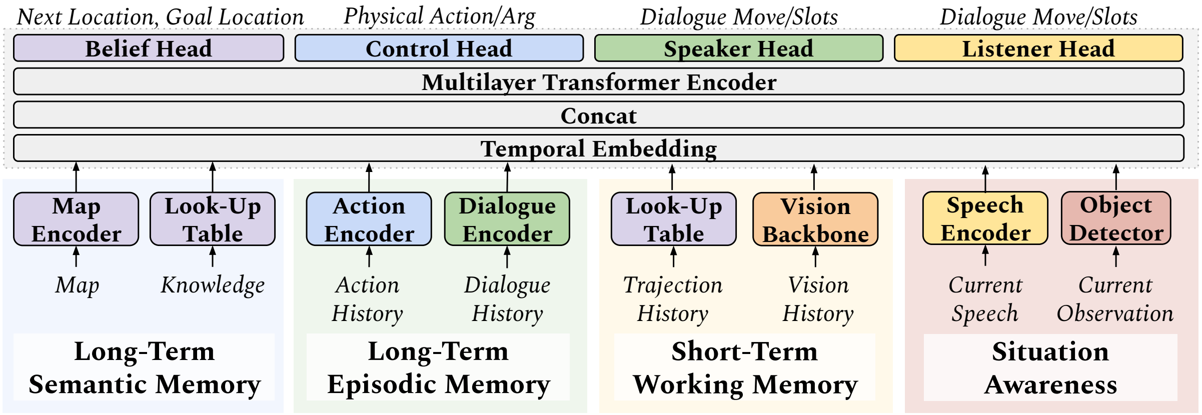


Figure 6: An overview of the architecture of the Temporally-Ordered Task-Oriented (TOTO) Transformer.

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.

Model	UfN (Seen)		RfN (Seen)		NfD (Seen)	
	Move Acc.	Slot F1	Move Acc.	Slot F1	Action Acc.	Act-Arg Joint Acc.
TOTO	40.9$_{(\pm 3.9)}$	36.9$_{(\pm 0.0)}$	29.2$_{(\pm 0.7)}$	55.7$_{(\pm 0.2)}$	41.2$_{(\pm 2.5)}$	36.0$_{(\pm 3.4)}$
TOTO (+ Belief Tracking)	39.5$_{(\pm 2.2)}$	37.0$_{(\pm 0.1)}$	28.8$_{(\pm 0.9)}$	55.7$_{(\pm 0.2)}$	40.7$_{(\pm 3.6)}$	34.0$_{(\pm 4.7)}$
TOTO (- Action History)	30.5$_{(\pm 1.5)}$	36.9$_{(\pm 0.0)}$	23.5$_{(\pm 1.7)}$	55.7$_{(\pm 0.0)}$	27.6$_{(\pm 2.8)}$	24.6$_{(\pm 4.0)}$
TOTO (- GT Transcript)	39.8$_{(\pm 1.9)}$	36.9$_{(\pm 0.1)}$	29.2$_{(\pm 0.8)}$	55.6$_{(\pm 0.1)}$	40.4$_{(\pm 3.4)}$	31.6$_{(\pm 4.3)}$
TOTO (- Object Detection)	42.5$_{(\pm 2.8)}$	37.0$_{(\pm 0.2)}$	30.4$_{(\pm 0.7)}$	55.8$_{(\pm 0.1)}$	39.2$_{(\pm 3.5)}$	34.4$_{(\pm 5.8)}$
TOTO (- Vision History)	41.9$_{(\pm 1.3)}$	37.0$_{(\pm 0.2)}$	29.1$_{(\pm 0.5)}$	55.8$_{(\pm 0.2)}$	42.0$_{(\pm 3.1)}$	36.1$_{(\pm 4.0)}$
TOTO (- Current Speech)	35.1$_{(\pm 2.7)}$	36.7$_{(\pm 0.5)}$	29.9$_{(\pm 0.9)}$	55.9$_{(\pm 0.2)}$	39.7$_{(\pm 1.9)}$	33.7$_{(\pm 3.0)}$
TOTO (- Map Knowledge)	42.6$_{(\pm 1.2)}$	36.9$_{(\pm 0.0)}$	29.3$_{(\pm 0.9)}$	55.8$_{(\pm 0.2)}$	44.6$_{(\pm 3.3)}$	39.1$_{(\pm 3.3)}$
Episodic Transformer	36.6$_{(\pm 3.6)}$	37.0$_{(\pm 0.2)}$	29.4$_{(\pm 1.2)}$	55.9$_{(\pm 0.2)}$	40.0$_{(\pm 2.8)}$	32.2$_{(\pm 4.0)}$
Fine-tuned BERT	**66.8**$_{(\pm \mathbf{2.0})}$	24.9$_{(\pm 5.5)}$	**52.7**$_{(\pm \mathbf{1.0})}$	46.0$_{(\pm 2.5)}$	32.4$_{(\pm 1.2)}$	16.2$_{(\pm 2.7)}$
Model	**UfN (Unseen)**		**RfN (Unseen)**		**NfD (Unseen)**	
	Move Acc.	**Slot F1**	**Move Acc.**	**Slot F1**	**Action Acc.**	**Act-Arg Joint Acc.**
TOTO	49.2$_{(\pm 3.0)}$	26.2$_{(\pm 0.0)}$	31.0$_{(\pm 1.7)}$	54.0$_{(\pm 0.7)}$	45.8$_{(\pm 3.8)}$	41.1$_{(\pm 2.8)}$
TOTO (+ Belief Tracking)	47.1$_{(\pm 3.5)}$	26.2$_{(\pm 0.0)}$	29.0$_{(\pm 2.0)}$	53.7$_{(\pm 0.7)}$	47.6$_{(\pm 4.5)}$	38.8$_{(\pm 3.1)}$
TOTO (- Action History)	35.5$_{(\pm 3.2)}$	26.1$_{(\pm 0.1)}$	28.2$_{(\pm 3.9)}$	54.8$_{(\pm 0.0)}$	36.8$_{(\pm 0.8)}$	36.0$_{(\pm 1.7)}$
TOTO (- GT Transcript)	46.7$_{(\pm 2.4)}$	26.2$_{(\pm 0.0)}$	31.6$_{(\pm 2.6)}$	54.2$_{(\pm 0.8)}$	46.2$_{(\pm 5.9)}$	37.6$_{(\pm 6.9)}$
TOTO (- Object Detection)	50.0$_{(\pm 1.8)}$	26.2$_{(\pm 0.1)}$	32.7$_{(\pm 2.2)}$	53.8$_{(\pm 1.2)}$	45.7$_{(\pm 5.2)}$	40.3$_{(\pm 5.4)}$
TOTO (- Vision History)	48.7$_{(\pm 2.3)}$	26.2$_{(\pm 0.1)}$	31.5$_{(\pm 2.9)}$	54.3$_{(\pm 0.7)}$	45.9$_{(\pm 4.2)}$	42.3$_{(\pm 3.5)}$
TOTO (- Current Speech)	42.8$_{(\pm 2.5)}$	25.8$_{(\pm 0.3)}$	33.8$_{(\pm 1.4)}$	55.1$_{(\pm 0.4)}$	46.5$_{(\pm 4.9)}$	39.4$_{(\pm 5.2)}$
TOTO (- Map Knowledge)	48.2$_{(\pm 1.0)}$	26.2$_{(\pm 0.1)}$	31.9$_{(\pm 1.2)}$	54.9$_{(\pm 0.8)}$	51.7$_{(\pm 3.4)}$	46.0$_{(\pm 4.0)}$
Episodic Transformer	45.1$_{(\pm 3.8)}$	26.1$_{(\pm 0.1)}$	33.4$_{(\pm 2.2)}$	54.7$_{(\pm 0.8)}$	46.6$_{(\pm 3.3)}$	37.0$_{(\pm 5.9)}$
Fine-tuned BERT	**67.2**$_{(\pm 1.5)}$	16.2$_{(\pm 3.5)}$	**57.0**$_{(\pm \mathbf{0.9})}$	46.9$_{(\pm 2.2)}$	37.1$_{(\pm 1.5)}$	19.6$_{(\pm 3.6)}$

Table 3: Experiment results of TOTO and baselines on the three tasks in the SDN benchmark. We use accuracy as the primary evaluation metric for navigation actions and dialogue moves, and use F1-score as the primary metric for dialogue slot, both in percentage. Each experiment is repeated with 5 random seeds. In each run, the model is validated on the complete validation fold (including both seen and unseen splits). The model is trained until overfit or reaching the max epochs. The model with lowest loss on the validation fold will be used for inference.

Ablation on Belief Tracking Prior work (Ma et al., 2019; Zhang and Chai, 2021) has indicated the power of task monitoring. To understand if end-to-end belief tracking would benefit the computational model, we additionally introduce a belief head with auxiliary loss, which is tasked to predict the location of the next timestamp on the trajectory and the location of the goal landmark annotated during data collection. According to Table 3, we observe marginal but no statistically significant improvement. This observation suggests that for long-range navigation tasks with unexpected goal changes, end-to-end approaches can hardly benefit from end-to-end belief tracking. Other modeling approaches should be explored to make full use of the rich belief update annotations of SDN.

8 Conclusion

We introduced DOROTHIE, 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 DOROTHIE 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 DOROTHIE 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, 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. A complete ethical statement is available in Appendix D.

References

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

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 J. Corso. 2020. The robotslang benchmark: Dialog-guided robot localization and navigation. In *Proceedings of the 4th Annual Conference on Robot Learning*.

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline C. Kowtko, Gwyneth Doherty-Sneddon, and Anne H. Anderson. 1997. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31.

Joyce Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. 2018. Language to action: Towards interactive task learning with physical agents. In *IJCAI*, pages 2–9.

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.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. 2021. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. 2018. End-to-end driving via conditional imitation learning. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 4693–4700. IEEE.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of oz studies-why and how. *Knowledge-based systems*, 6(4):258–266.

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. *arXiv preprint arXiv:1807.03367*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16.

Marius Dupuis and Han Grezlikowski. 2006. Opendrive®-an open standard for the description of roads in driving simulations. In *Proceedings of the Driving Simulation Conference*, pages 25–36.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

Uwe Franke. 2017. Autonomous driving. *Computer Vision in Vehicle Technology: Land, Sea & Air*, pages 24–54.

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 (RA-L)*.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Jing Gu, Eliana Stefani, Qi Wu, Jesse Thomason, and Xin Eric Wang. 2022. Vision-and-language navigation: A survey of tasks, methods, and future directions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

John HL Hansen, Xianxian Zhang, Murat Akbacak, Umit H Yapanel, Bryan Pellom, Wayne Ward, and Pongtep Angkititrakul. 2005. Cu-move: Advanced in-vehicle speech systems for route navigation. In *DSP for in-vehicle and mobile systems*, pages 19–45. Springer.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Karl Moritz Hermann, Mateusz Malinowski, Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, and Raia Hadsell. 2020. Learning to follow directions in street view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11773–11781.

Yicong Hong, Zun Wang, Qi Wu, and Stephen Gould. 2022. Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Thomas Victor Ilyevsky, Jared Sigurd Johansen, and Jeffrey Mark Siskind. 2021. Talk the talk and walk the walk: Dialogue-driven navigation in unknown indoor environments. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4903–4910. IEEE.

Muhammad Zubair Irshad, Chih-Yao Ma, and Zsolt Kira. 2021. Hierarchical cross-modal agent for robotics vision-and-language navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13238–13246. IEEE.

Nobuo Kawaguchi, Shigeki Matsubara, Yukiko Yamaguchi, Kazuya Takeda, and Fumitada Itakura. 2004. Ciair in-car speech database. In *Eighth International Conference on Spoken Language Processing*.

David Kirsh and Paul Maglio. 1994. On distinguishing epistemic from pragmatic action. *Cognitive science*, 18(4):513–549.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. 2020. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *European Conference on Computer Vision*, pages 104–120. Springer.

Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. 2016. Professor forcing: A new algorithm for training recurrent networks. *Advances in neural information processing systems*, 29.

Bingqian Lin, Yi Zhu, Yanxin Long, Xiaodan Liang, Qixiang Ye, and Liang Lin. 2021. Adversarial reinforced instruction attacker for robust vision-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. 2019. Self-monitoring navigation agent via auxiliary progress estimation. In *International Conference on Learning Representations*.

Yi Ma, Antoine Raux, Deepak Ramachandran, and Rakesh Gupta. 2012. Landmark-based location belief tracking in a spoken dialog system. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 169–178.

Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. 2006. Walk the talk: Connecting language, knowledge, and action in route instructions. *Def*, 2(6):4.

Matthew Marge, Claire Bonial, Ashley Foots, Cory Hayes, Cassidy Henry, Kimberly Pollard, Ron Artstein, Clare Voss, and David Traum. 2017. Exploring variation of natural human commands to a robot in a collaborative navigation task. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 58–66, Vancouver, Canada.

Matthew Marge, Felix Gervits, Gordon Briggs, Matthias Scheutz, and Antonio Roque. 2020. Let's do that first! a comparative analysis of instruction-giving in human-human and human-robot situated dialogue. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*.

Teruhisa Misu, Antoine Raux, Rakesh Gupta, and Ian Lane. 2014. Situated language understanding at 25 miles per hour. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 22–31.

Matthias Mueller, Alexey Dosovitskiy, Bernard Ghanem, and Vladlen Koltun. 2018. Driving policy transfer via modularity and abstraction. In *Proceedings of The 2nd Conference on Robot Learning*, volume 87 of *Proceedings of Machine Learning Research*, pages 1–15.

Khanh Nguyen and Hal Daumé III. 2019. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. 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 684–695, Hong Kong, China.

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 *AAAI*.

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

Tzuf Paz-Argaman and Reut Tsarfaty. 2019. RUN through the streets: A new dataset and baseline models for realistic urban navigation. 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 6449–6455, Hong Kong, China. Association for Computational Linguistics.

Deepak Ramachandran, Igor V Karpov, Rakesh Gupta, and Antoine Raux. 2013. Driver familiarity modeling for generating navigation directions. In *16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013)*, pages 2193–2200. IEEE.

Laurel D Riek. 2012. Wizard of oz studies in hri: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, 1(1):119–136.

Junha Roh, Chris Paxton, Andrzej Pronobis, Ali Farhadi, and Dieter Fox. 2020. Conditional driving from natural language instructions. In *Proceedings of the Conference on Robot Learning*, pages 540–551.

Francisca Rosique, Pedro J Navarro, Carlos Fernández, and Antonio Padilla. 2019. A systematic review of perception system and simulators for autonomous vehicles research. *Sensors*, 19(3):648.

Raphael Schumann and Stefan Riezler. 2021. Generating landmark navigation instructions from maps as a graph-to-text problem. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 489–502, Online. Association for Computational Linguistics.

Lanbo She and Joyce Chai. 2017. Interactive learning of grounded verb semantics towards human-robot communication. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1634–1644.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10740–10749.

J.M.H. Sinclair, J.M.H. Sinclair, M. Coulthard, R.M. Coulthard, and Social Science Research Council (Great Britain). 1975. *Towards an Analysis of Discourse: The English Used by Teachers and Pupils*. Oxford University Press.

John Skinner, Sourav Garg, Niko Sünderhauf, Peter Corke, Ben Upcroft, and Michael Milford. 2016. High-fidelity simulation for evaluating robotic vision performance. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2737–2744. IEEE.

Dimitris Spiliotopoulos, Ion Androutsopoulos, and Constantine D Spyropoulos. 2001. Human-robot interaction based on spoken natural language dialogue. In *Proceedings of the European workshop on service and humanoid robots*, pages 25–27. Citeseer.

NN Sriram, Tirth Maniar, Jayaganesh Kalyanasundaram, Vineet Gandhi, Brojeshwar Bhowmick, and K Madhava Krishna. 2019. Talk to the vehicle: Language conditioned autonomous navigation of self driving cars. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5284–5290. IEEE.

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

David Traum, Cassidy Henry, Stephanie Lukin, Ron Artstein, Felix Gervits, Kimberly Pollard, Claire Bonial, Su Lei, Clare Voss, Matthew Marge, et al. 2018. Dialogue structure annotation for multi-floor interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. 2021. Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory. *International Journal of Computer Vision*, 129(1):246–266.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations (ICLR)*.

Fuliang Weng, Pongtep Angkititrakul, Elizabeth E Shriberg, Larry Heck, Stanley Peters, and John HL Hansen. 2016. Conversational in-vehicle dialog systems: The past, present, and future. *IEEE Signal Processing Magazine*, 33(6):49–60.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. 2021. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34.

Yichi Zhang and Joyce Chai. 2021. Hierarchical task learning from language instructions with unified transformers and self-monitoring. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4202–4213.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2021. Deformable {detr}: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*.

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 (Riek, 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, DOROTHIE 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 (DOROTHIE) 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* ϕ, *steering angle* θ, 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* $\{\texttt{str}_i\}$ and (possibly incomplete) *landmarks* $\{\texttt{lm}_i\}$ on the *map* topology M. At any time t, the *interaction history* H_t includes previous *observation* and *dialogue* $\{O^H_{t-1}, \langle U^H_{t-1,\text{hum}}, U^H_{t-1,\text{agent}}\rangle\}$. 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.

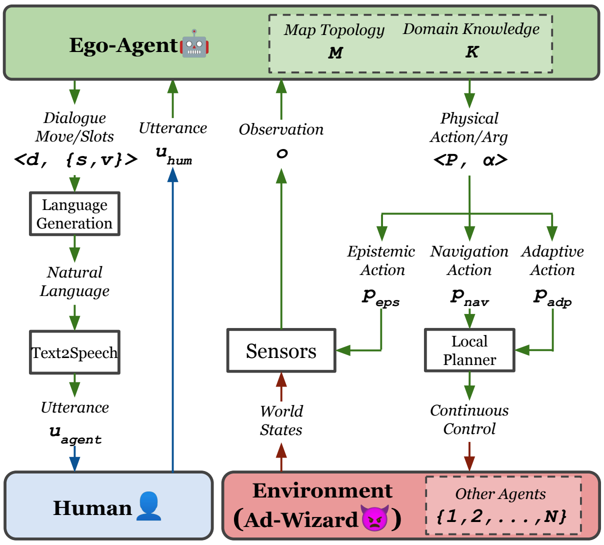


Figure 7: An overview of the framework of the DOROTHIE platform.

Navigation Actions To navigate, the agent has a discrete and finite physical action space $\mathcal{P} = \mathcal{P}_{\text{nav}} \cup \mathcal{P}_{\text{adp}} \cup \mathcal{P}_{\text{eps}}$. Each *navigation action* in $\mathcal{P}_{\text{nav}} = \{\texttt{LaneFollow}, \texttt{LaneSwitch}, \texttt{JTurn}, \texttt{UTurn}, \texttt{Stop}, \texttt{Start}\}$ 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 $\pi \in \Pi : \mathcal{P}_{\text{nav}} \times \Theta \to \mathcal{A}$.

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

Sensors and Epistemic Action The sensors of the agent defines an observation function $\omega \in \Omega : \mathcal{S} \to \mathcal{O}$ 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 function $\mathcal{P}_{\text{eps}} : \Omega \to \Omega$. We assume for now a fixed first-person RGB camera with no noise, with only $\mathcal{P}_{\text{eps}} = \{\texttt{LightChange}\}$.

Interaction To interact with human, the agent select a the dialogue move-slots pair $\langle d, \{s, v\}\rangle$, as described in Section 4.1. The language generation module generates natural language as a sequence of tokens and produces the utterance u_{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:

```
{
  "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.",

  "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"}
  ],

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

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

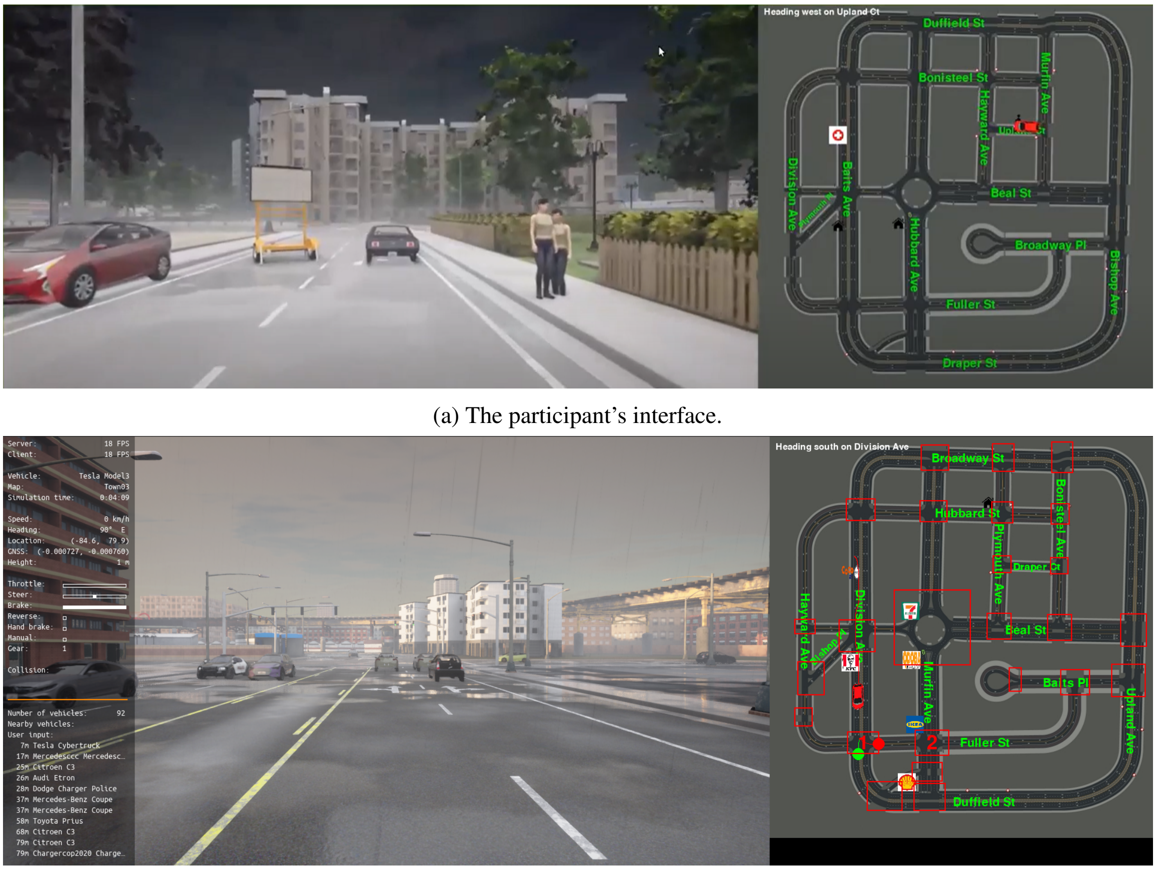
Knowledge Disparity Both the Co-Wizard and the participant perceive the environment through a stream of RGB images. To replicate realistic outdoor navigation, an aerial map of the environment, with landmarks and current location, is provided to both the participant and the Co-Wizard. While both players 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.

```
{
  "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/



(a) The participant's interface.

```
Server:                 18 FPS
Client:                 18 FPS

Vehicle:           Tesla Model3
Map:                     Town03
Simulation time:        0:04:09

Speed:                   0 km/h
Heading:                 90°  E
Location:          (-84.6, 79.9)
GNSS:  (-0.000727, -0.000760)
Height:                     1 m

Throttle:
Steer:
Brake:
Reverse:        □
Hand brake:     □
Manual:         □
Gear:           1

Collision:

Number of vehicles:        92
Nearby vehicles:
User input:
   7m Tesla Cybertruck
  17m Mercedesccc Mercedesc…
  25m Citroen C3
  26m Audi Etron
  28m Dodge Charger Police
  37m Mercedes-Benz Coupe
  37m Mercedes-Benz Coupe
  58m Toyota Prius
  68m Citroen C3
  79m Citroen C3
  79m Chargercop2020 Charge…
```

(b) The Co-Wizard's interface.

Figure 8: Graphical interfaces for the participant and Co-Wizard.

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.

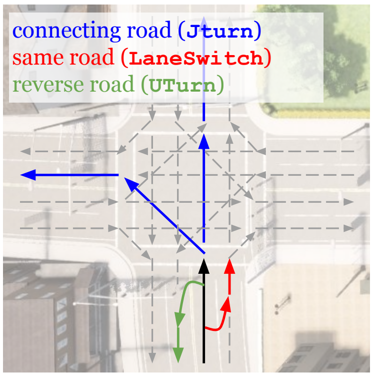
Aerial View Both the participant and the Co-Wizard have access to a 2D aerial map of the town, with the location and heading direction of the ego-vehicle. Various annotations are included to facilitate operation, including trajectory history, street names and landmarks. The Co-Wizard is additionally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

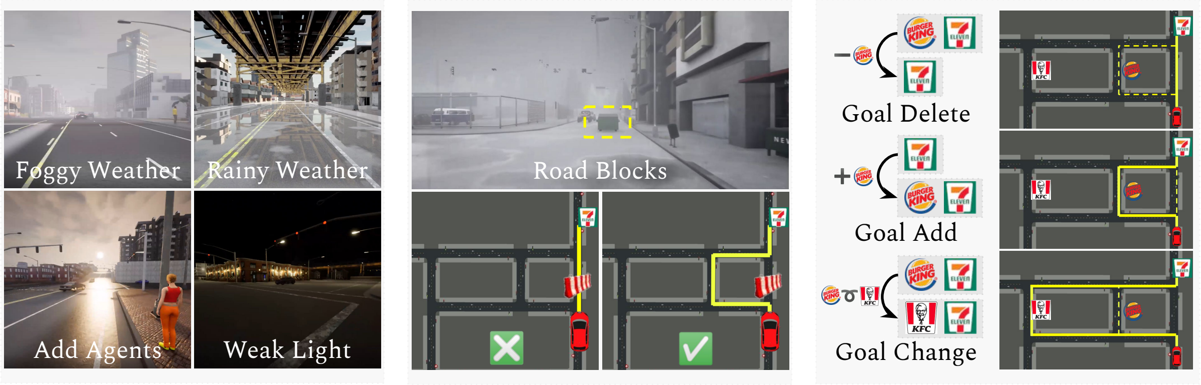
A.6 Co-Wizard Interface and Actions

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



Physical Actions	Args	Descriptions
`LaneFollow`	-	Default behaviour, follow the current lane.
`LaneSwitch`	Angle (Rotation)	Switch to a neighboring lane.
`JTurn`	Angle (Rotation)	Turn to a connecting road at a junction.
`UTurn`	-	Make a U-turn to the opposite direction.
`Stop`	-	Brake the vehicle manually.
`Start`	-	Start the vehicle manually.
`SpeedChange`	Speed (± 5)	Change the desired cruise speed by 5 km/h.
`LightChange`	Light State (`On`/`Off`)	Change the front light state.
Mental Actions	**Args**	**Descriptions**
`PlanUpdate`	`List[Junction ID]`	Indicate intended trajectory towards a destination.
`GoalUpdate`	`List[Landmark]`	Indicate current goal as an intended landmark.
`StatusUpdate`	`Tuple[Landmark,Status]`	Indicate a change in task status.
`KnowledgeUpdate`	`x,y`	Guess the location of an unknown landmark.
`Other`	-	Other belief state updates.

Figure 9: The space of primitive physical actions and mental actions of *Co-Wizard*.



(a) Environmental Changes.

(b) Unexpected event at plan level.

(c) Unexpected event at goal level.

Figure 10: The Ad-Wizard is able to change the environment and tasks on-the-fly.

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[3] 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

[3] https://pypi.org/project/SpeechRecognition/

initiated and ends when it is completed or abandoned. We observed in the corpus that the participant sometimes describes the next tasks to perform when current task is still ongoing, leading to small pieces of conversation standing alone from the major transaction unit for that subgoal. Due to the addition, change, or deletion of a subgoal, some transactions are interrupted and continued afterwards (Figure 4). Therefore, we assign each utterance to the subgoal it aims for, and one of the task status in `Ongoing`, `Complete`, `Abandoned`, `Pending` to each subgoal whenever there is a change specified by the utterance.

Annotating Exchanges *Exchange Units (EUs)* are sequences of dialogue moves towards common ground. They starts 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 overlaps 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 $\kappa = 0.77 \pm 0.02$ for dialogue move annotation, and $\kappa = 0.85 \pm 0.02$ 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 let's go to Baits 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
[Bot] I'll make a u-turn and then go to the Division Place
[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
[Bot] 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

[Bot] I think we're at the KFC now

[Human] okay great now we're going to go back to my aunt's house

[Bot] okay

[Human] 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

[Human] do you remember where it is

[Bot] yes I do

[Human] what's the location of my aunt's house

[Bot] 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

[Human] that's correct that's where we're going right now

[Bot] okay I'll take you there

[Human] and once we get there we're just going to park on the side of the road

[Bot] 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 a embedding with a size of 300 × 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: 6×10^{-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: 2×10^{-4}
- backbone learning rate: 2×10^{-5}
- weight decay: 1×10^{-4}
- optimizer: AdamW

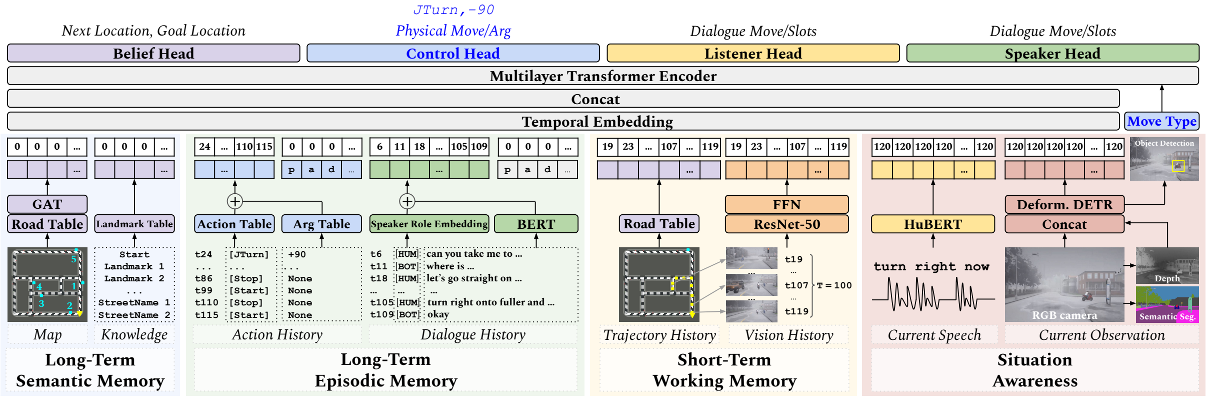


Figure 11: Temporally-Ordered Task-Oriented (TOTO) Transformer.

Learning Parameters

- learning rate: 1×10^{-5}
- weight decay: 1×10^{-4}
- optimizer: AdamW

Loss coefficients

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

Semantic Segmentation			Object Detection	
mIoU	mAcc	aAcc	$\text{AP}_{(\text{IoU}=0.50)}$	$\text{AP}_{(\text{IoU}=0.50:0.95)}$
65.60%	71.60%	97.10%	71.40%	41.30%

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

Additional Ablations We provide complete ablation results in Table 5.

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

Model	**UfN (Seen)**		**RfN (Seen)**		**NfD (Seen)**	
	Move Acc.	**Slot F1**	**Move Acc.**	**Slot F1**	**Action Acc.**	**Act-Arg Joint Acc.**
TOTO	40.9$_{(\pm 3.9)}$	36.9$_{(\pm 0.0)}$	29.2$_{(\pm 0.7)}$	55.7$_{(\pm 0.2)}$	41.2$_{(\pm 2.5)}$	36.0$_{(\pm 3.4)}$
TOTO (+ Belief Tracking)	39.5$_{(\pm 2.2)}$	37.0$_{(\pm 0.1)}$	28.8$_{(\pm 0.9)}$	55.7$_{(\pm 0.2)}$	40.7$_{(\pm 3.6)}$	34.0$_{(\pm 4.7)}$
TOTO (+ Fine-tuned BERT)	38.4$_{(\pm 2.7)}$	36.9$_{(\pm 0.0)}$	27.8$_{(\pm 0.7)}$	55.7$_{(\pm 0.0)}$	43.6$_{(\pm 1.6)}$	30.0$_{(\pm 4.1)}$
TOTO (- Action History)	30.5$_{(\pm 1.5)}$	36.9$_{(\pm 0.0)}$	23.5$_{(\pm 1.7)}$	55.7$_{(\pm 0.0)}$	27.6$_{(\pm 2.8)}$	24.6$_{(\pm 4.0)}$
TOTO (- GT Transcript)	39.8$_{(\pm 1.9)}$	36.9$_{(\pm 0.1)}$	29.2$_{(\pm 0.8)}$	55.6$_{(\pm 0.1)}$	40.4$_{(\pm 3.4)}$	31.6$_{(\pm 4.3)}$
TOTO (- Object Detection)	42.5$_{(\pm 2.8)}$	37.0$_{(\pm 0.2)}$	30.4$_{(\pm 0.7)}$	55.8$_{(\pm 0.1)}$	39.2$_{(\pm 3.5)}$	34.4$_{(\pm 5.8)}$
TOTO (- Vision History)	41.9$_{(\pm 1.3)}$	37.0$_{(\pm 0.2)}$	29.1$_{(\pm 0.5)}$	55.8$_{(\pm 0.2)}$	42.0$_{(\pm 3.1)}$	36.1$_{(\pm 4.0)}$
TOTO (- Current Speech)	35.1$_{(\pm 2.7)}$	36.7$_{(\pm 0.5)}$	29.9$_{(\pm 0.9)}$	55.9$_{(\pm 0.2)}$	39.7$_{(\pm 1.9)}$	33.7$_{(\pm 3.0)}$
TOTO (- Map Knowledge)	42.6$_{(\pm 1.2)}$	36.9$_{(\pm 0.0)}$	29.3$_{(\pm 0.9)}$	55.8$_{(\pm 0.2)}$	44.6$_{(\pm 3.3)}$	39.1$_{(\pm 3.3)}$
Fine-tuned BERT Only	66.8$_{(\pm 2.0)}$	24.9$_{(\pm 5.5)}$	52.7$_{(\pm 1.0)}$	46.0$_{(\pm 2.5)}$	32.4$_{(\pm 1.2)}$	16.2$_{(\pm 2.7)}$
BERT Only	52.1$_{(\pm 3.2)}$	39.9$_{(\pm 1.3)}$	52.3$_{(\pm 1.0)}$	56.1$_{(\pm 0.3)}$	30.4$_{(\pm 1.8)}$	25.6$_{(\pm 2.8)}$
HuBERT Only	35.7$_{(\pm 3.4)}$	36.9$_{(\pm 0.0)}$	24.5$_{(\pm 0.1)}$	55.6$_{(\pm 0.1)}$	31.2$_{(\pm 0.9)}$	31.1$_{(\pm 0.9)}$
Deformable-DETR Only	31.5$_{(\pm 0.0)}$	36.9$_{(\pm 0.0)}$	24.4$_{(\pm 0.0)}$	55.7$_{(\pm 0.0)}$	31.5$_{(\pm 0.6)}$	31.5$_{(\pm 0.6)}$
Map Encoder Only	25.3$_{(\pm 2.5)}$	36.9$_{(\pm 0.0)}$	22.5$_{(\pm 1.6)}$	55.7$_{(\pm 0.0)}$	29.7$_{(\pm 0.6)}$	28.4$_{(\pm 1.8)}$
E.T.	36.6$_{(\pm 3.6)}$	37.0$_{(\pm 0.2)}$	29.4$_{(\pm 1.2)}$	55.9$_{(\pm 0.2)}$	40.0$_{(\pm 2.8)}$	32.2$_{(\pm 4.0)}$
E.T. (+ Fine-tuned BERT)	33.6$_{(\pm 1.1)}$	36.8$_{(\pm 0.1)}$	26.5$_{(\pm 0.9)}$	55.7$_{(\pm 0.0)}$	38.0$_{(\pm 1.1)}$	27.6$_{(\pm 6.2)}$
Model	**UfN (Unseen)**		**RfN (Unseen)**		**NfD (Unseen)**	
	Move Acc.	**Slot F1**	**Move Acc.**	**Slot F1**	**Action Acc.**	**Act-Arg Joint Acc.**
TOTO	49.2$_{(\pm 3.0)}$	26.2$_{(\pm 0.0)}$	31.0$_{(\pm 1.7)}$	54.0$_{(\pm 0.7)}$	45.8$_{(\pm 3.8)}$	41.1$_{(\pm 2.8)}$
TOTO (+ Belief Tracking)	47.1$_{(\pm 3.5)}$	26.2$_{(\pm 0.0)}$	29.0$_{(\pm 2.0)}$	53.7$_{(\pm 0.7)}$	47.6$_{(\pm 4.5)}$	38.8$_{(\pm 3.1)}$
TOTO (+ Fine-tuned BERT)	49.6$_{(\pm 0.9)}$	26.2$_{(\pm 0.1)}$	34.0$_{(\pm 2.1)}$	54.8$_{(\pm 0.0)}$	48.5$_{(\pm 4.6)}$	36.2$_{(\pm 5.5)}$
TOTO (- Action History)	35.5$_{(\pm 3.2)}$	26.1$_{(\pm 0.1)}$	28.2$_{(\pm 3.9)}$	54.8$_{(\pm 0.0)}$	36.8$_{(\pm 0.8)}$	36.0$_{(\pm 1.7)}$
TOTO (- GT Transcript)	46.7$_{(\pm 2.4)}$	26.2$_{(\pm 0.0)}$	31.6$_{(\pm 2.6)}$	54.2$_{(\pm 0.8)}$	46.2$_{(\pm 5.9)}$	37.6$_{(\pm 6.9)}$
TOTO (- Object Detection)	50.0$_{(\pm 1.8)}$	26.2$_{(\pm 0.1)}$	32.7$_{(\pm 2.2)}$	53.8$_{(\pm 1.2)}$	45.7$_{(\pm 5.2)}$	40.3$_{(\pm 5.4)}$
TOTO (- Vision History)	48.7$_{(\pm 2.3)}$	26.2$_{(\pm 0.1)}$	31.5$_{(\pm 2.9)}$	54.3$_{(\pm 0.7)}$	45.9$_{(\pm 4.2)}$	42.3$_{(\pm 3.5)}$
TOTO (- Current Speech)	42.8$_{(\pm 2.5)}$	25.8$_{(\pm 0.3)}$	33.8$_{(\pm 1.4)}$	55.1$_{(\pm 0.4)}$	46.5$_{(\pm 4.9)}$	39.4$_{(\pm 5.2)}$
TOTO (- Map Knowledge)	48.2$_{(\pm 1.0)}$	26.2$_{(\pm 0.1)}$	31.9$_{(\pm 1.2)}$	54.9$_{(\pm 0.8)}$	51.7$_{(\pm 3.4)}$	46.0$_{(\pm 4.0)}$
Fine-tuned BERT Only	67.2$_{(\pm 1.5)}$	16.2$_{(\pm 3.5)}$	57.0$_{(\pm 0.9)}$	46.9$_{(\pm 2.2)}$	37.1$_{(\pm 1.5)}$	19.6$_{(\pm 3.6)}$
BERT Only	57.3$_{(\pm 2.1)}$	31.7$_{(\pm 2.1)}$	57.4$_{(\pm 1.3)}$	55.9$_{(\pm 0.6)}$	35.3$_{(\pm 3.6)}$	30.1$_{(\pm 4.2)}$
HuBERT Only	40.9$_{(\pm 2.3)}$	26.2$_{(\pm 0.0)}$	30.2$_{(\pm 0.2)}$	54.7$_{(\pm 0.1)}$	36.7$_{(\pm 0.1)}$	36.7$_{(\pm 0.2)}$
Deformable-DETR Only	37.6$_{(\pm 0.0)}$	26.2$_{(\pm 0.0)}$	30.7$_{(\pm 0.0)}$	54.8$_{(\pm 0.0)}$	36.9$_{(\pm 0.1)}$	36.9$_{(\pm 0.1)}$
Map Encoder Only	30.1$_{(\pm 9.9)}$	26.2$_{(\pm 0.0)}$	23.1$_{(\pm 10.6)}$	54.8$_{(\pm 0.0)}$	37.0$_{(\pm 0.1)}$	37.0$_{(\pm 0.1)}$
E.T.	45.1$_{(\pm 3.8)}$	26.1$_{(\pm 0.1)}$	33.4$_{(\pm 2.2)}$	54.7$_{(\pm 0.8)}$	46.6$_{(\pm 3.3)}$	37.0$_{(\pm 5.9)}$
E.T. (+ Fine-tuned BERT)	41.8$_{(\pm 0.6)}$	26.2$_{(\pm 0.0)}$	34.1$_{(\pm 1.1)}$	54.8$_{(\pm 0.0)}$	44.5$_{(\pm 3.1)}$	32.9$_{(\pm 7.3)}$

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.