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

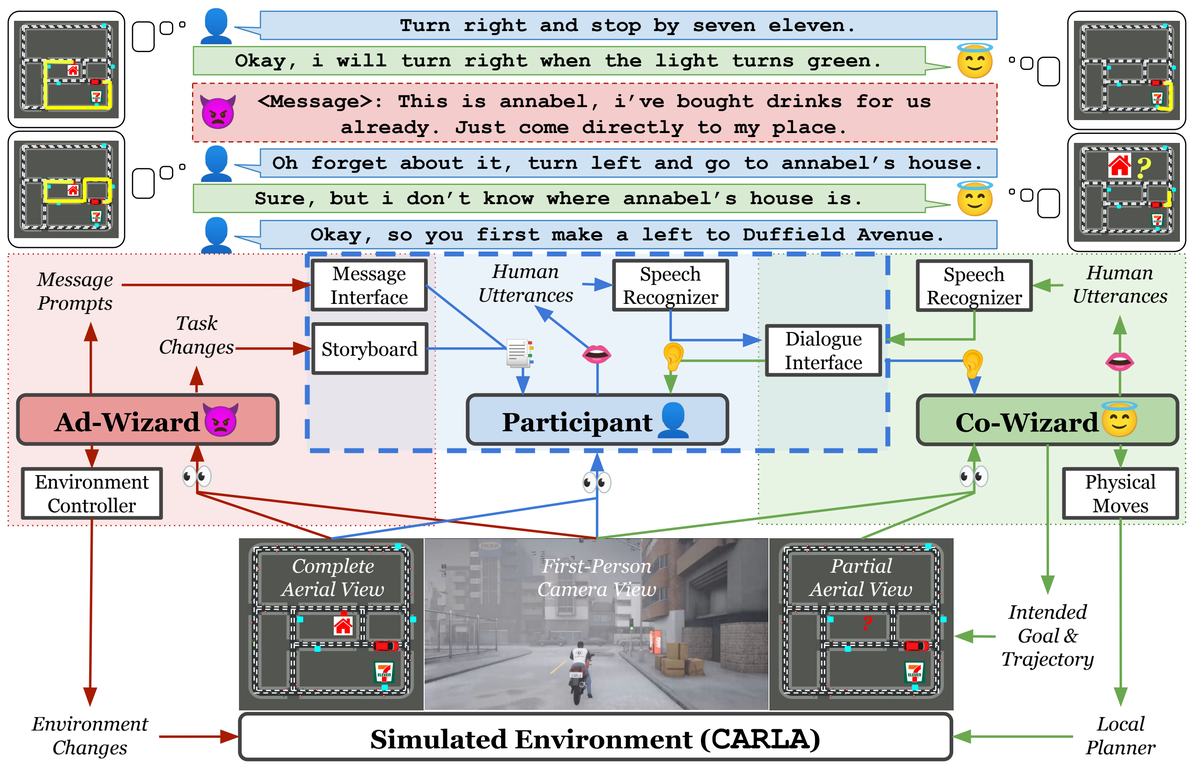
Ziqiao Ma, Ben VanDerPloeg*, Cristian-Paul Bara*, Yidong Huang*,
Eui-In Kim, Felix Gervits, Matthew Marge, Joyce Chai

Motivation

Current vision-and-language navigation (VLN) setups only reflect partial and simplified challenges compared to those faced by autonomous vehicles (AVs), which navigate in highly dynamic environment with continuous physical control. When unexpected situations arise, agents should collaborate with human operators in the form of spoken dialogue. It's thus important to empower AVs with the ability to harness human knowledge and expertise and to enable natural language communication and collaboration in tackling unexpected situations. In this work, we seek to enable AVs to navigate in **continuous** and **dynamic** environment, and communicate with human through **sensorimotor** grounded **dialogue**.

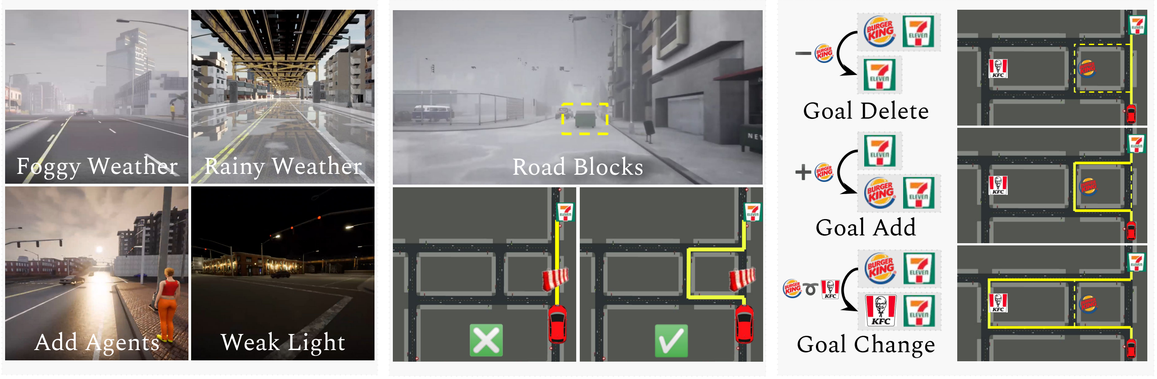
Environment and Human Studies

We developed a novel framework, **Dialogue On the ROad To Handle Irregular Events** (DOROTHIE) upon CARLA [1] to study situated human-vehicle communication based on the **Wizard-of-Oz (WoZ)** setting. We extend the traditional single-wizard to a duo-wizard setup approach by introducing a pair of Wizards.



- The **naïve participant** communicates with the vehicle to visit goal locations specified in a storyboard.
- The **Cooperative-Wizard** controls the agent's behaviors and carries language communication with the human participant to jointly achieve the goal.
- The **Adversarial-Wizard** controls the environment and task interface and introduces unexpected situation on-the-fly.

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.
1-3 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.



Dataset

We recruit 40 human subjects and collect Situated Dialogue Navigation (SDN), a fine-grained navigation benchmark of 183 trials.

We annotate each dialogue session with 4 levels of linguistic units.

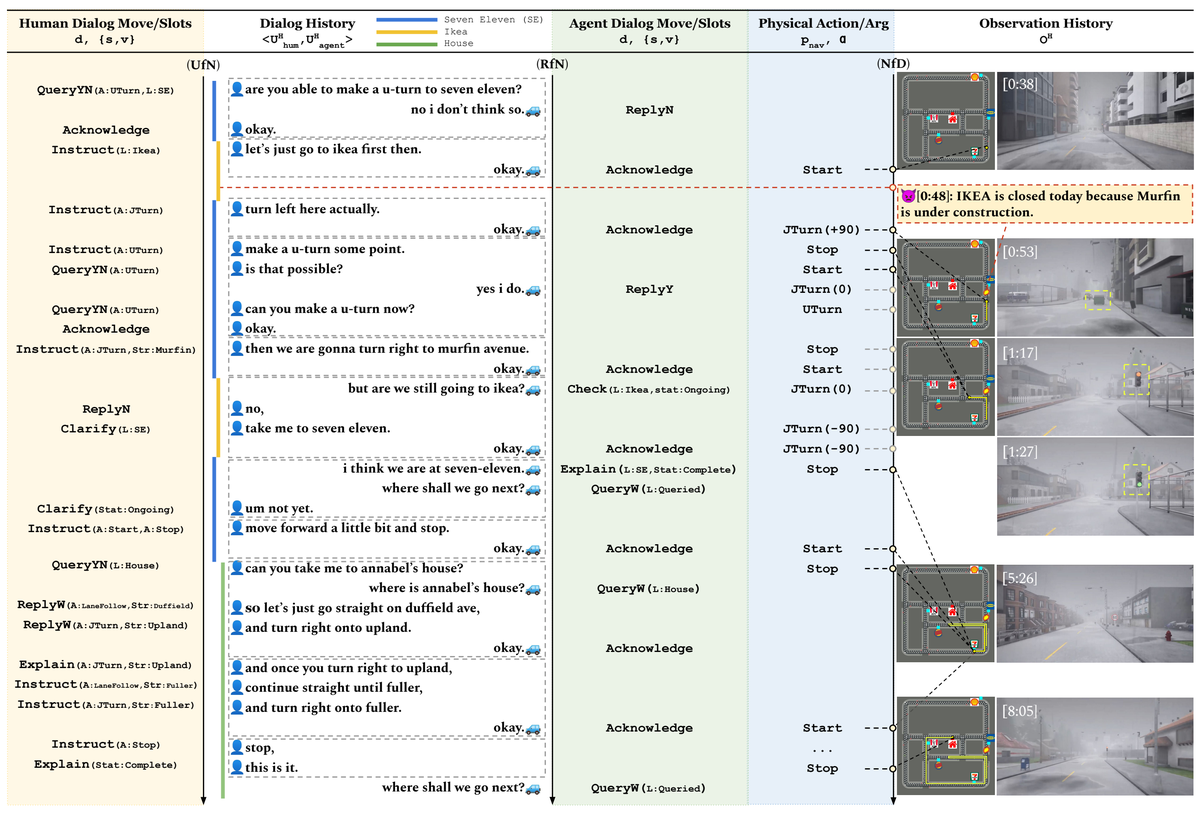
- Transaction Units (TUs)
- Exchange Units (EUs)
- Dialogue Moves
- Dialogue Slots

Metric	Value
Control Stream	18.7 h
Trimmed Audio	2.9 h
# Sessions	183
# Utterances	8415
# Words	50398
Vocabulary	1373
# Transactions	578
# Exchanges	4089
# Dialog Moves	11623
# Slot Values	8618
# Physical Actions	9448

Task Definition

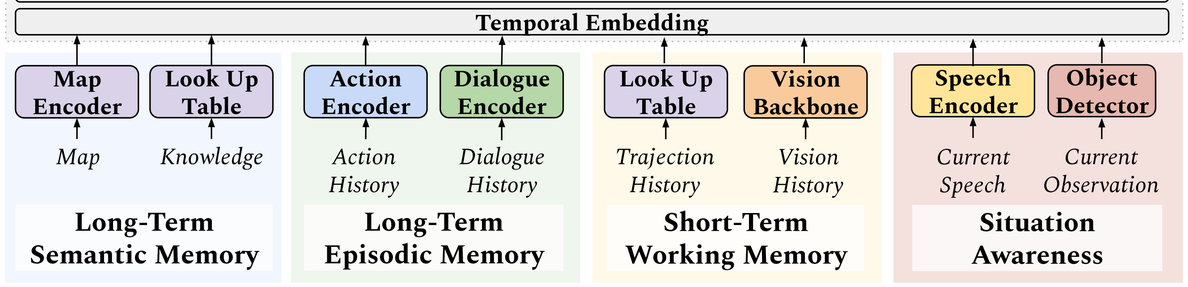
We evaluate the agent's ability in predicting dialogue moves from humans as well as generating its own dialogue moves and physical navigation actions.

- When: human speaks or agent selects a dialogue/navigation action
- Input: history of dialogue, RGB sensors, speech, and actions
- Output: human's current dialogue move/slot (**UfD**), agent's next dialogue move/slot (**RfD**), agent's next physical action (**NfD**)



Computational Model and Experiments

We present Temporally-Ordered Task-Oriented Transformer (TOTO), a Transformer-based baseline.



We summarize the experiment results in the table below.

Model	UfN (Seen) Move Acc.	UfN (Seen) Slot F1	RfN (Seen) Move Acc.	RfN (Seen) Slot F1	NfD (Seen) Action Acc.	NfD (Seen) 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	$66.8_{(\pm 2.0)}$	$24.9_{(\pm 5.5)}$	$52.7_{(\pm 1.0)}$	$46.0_{(\pm 2.5)}$	$32.4_{(\pm 1.2)}$	$16.2_{(\pm 2.7)}$
Model	**UfN (Unseen) Move Acc.**	**UfN (Unseen) Slot F1**	**RfN (Unseen) Move Acc.**	**RfN (Unseen) Slot F1**	**NfD (Unseen) Action Acc.**	**NfD (Unseen) 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	$67.2_{(\pm 1.5)}$	$16.2_{(\pm 3.5)}$	$57.0_{(\pm 0.9)}$	$46.9_{(\pm 2.2)}$	$37.1_{(\pm 1.5)}$	$19.6_{(\pm 3.6)}$

Overall, our preliminary experiment has shown that the tasks are challenging.

- TOTO 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;
- The finetuned language model only masters dialogue move tasks and the Episodic Transformer underperforms language tasks.
- Action history is crucial in the understanding and prediction of navigation/dialogue actions;

References

[1] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. CARLA: An open urban driving simulator. In Proceedings of the 1st Annual Conference on Robot Learning, pages 1–16, 201

Links