IROS '24
ABU DHABI

DriVLMe: Enhancing LLM-based Autonomous Driving Agents with Embodied and Social Experiences

Yidong Huang[1], Jacob Sansom[1], Ziqiao Ma[1], Felix Gervits[2], Joyce Chai[1] [1]University of Michigan; [2]ARL

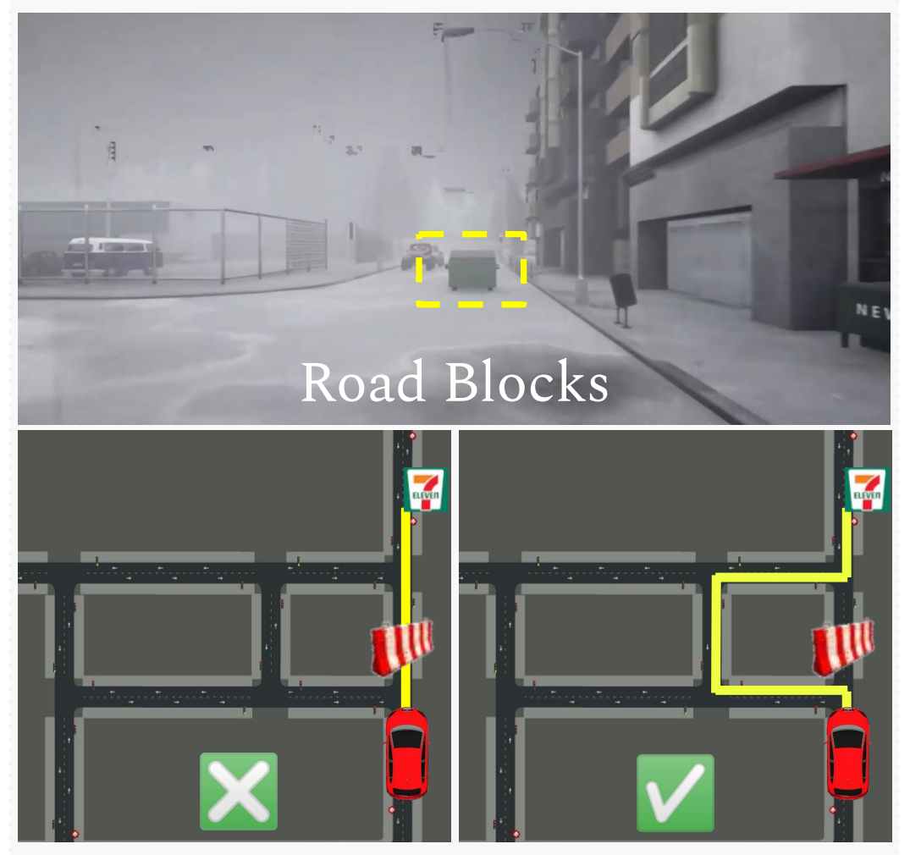
UNIVERSITY OF MICHIGAN
DEVCOM ARMY RESEARCH LABORATORY

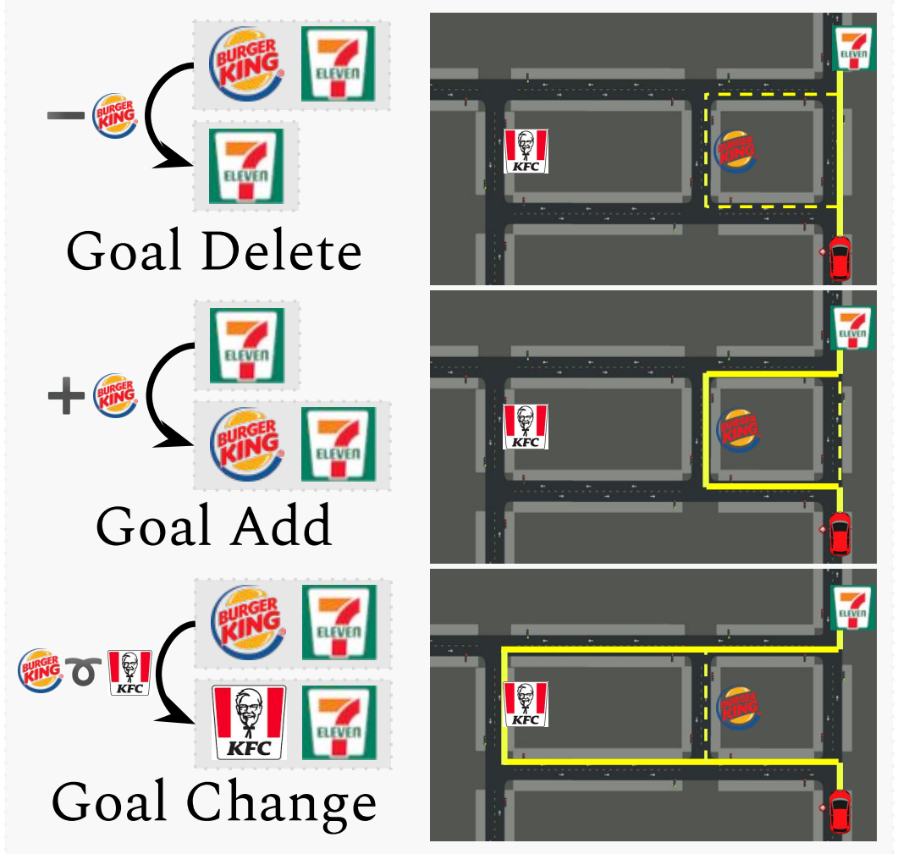
Motivation

An increasing number of efforts have demonstrated the potential of foundation models in the field of autonomous driving. However, the experimental setups are preliminary and simplified compared to the real driving scenarios in human environments faced by autonomous vehicles (AVs).

- Agents should be able to plan long-horizon navigations in highly dynamic environments;
- Agents should collaborate with humans in spoken dialogue in unexpected situations.

In this work, we study the capabilities and boundaries of LLM-based AV agents, which are developed from both **embodied** and **social** experiences, and tasked to navigate in **continuous** and **dynamic** environments and communicate with humans through **sensorimotor** grounded **dialogue**.

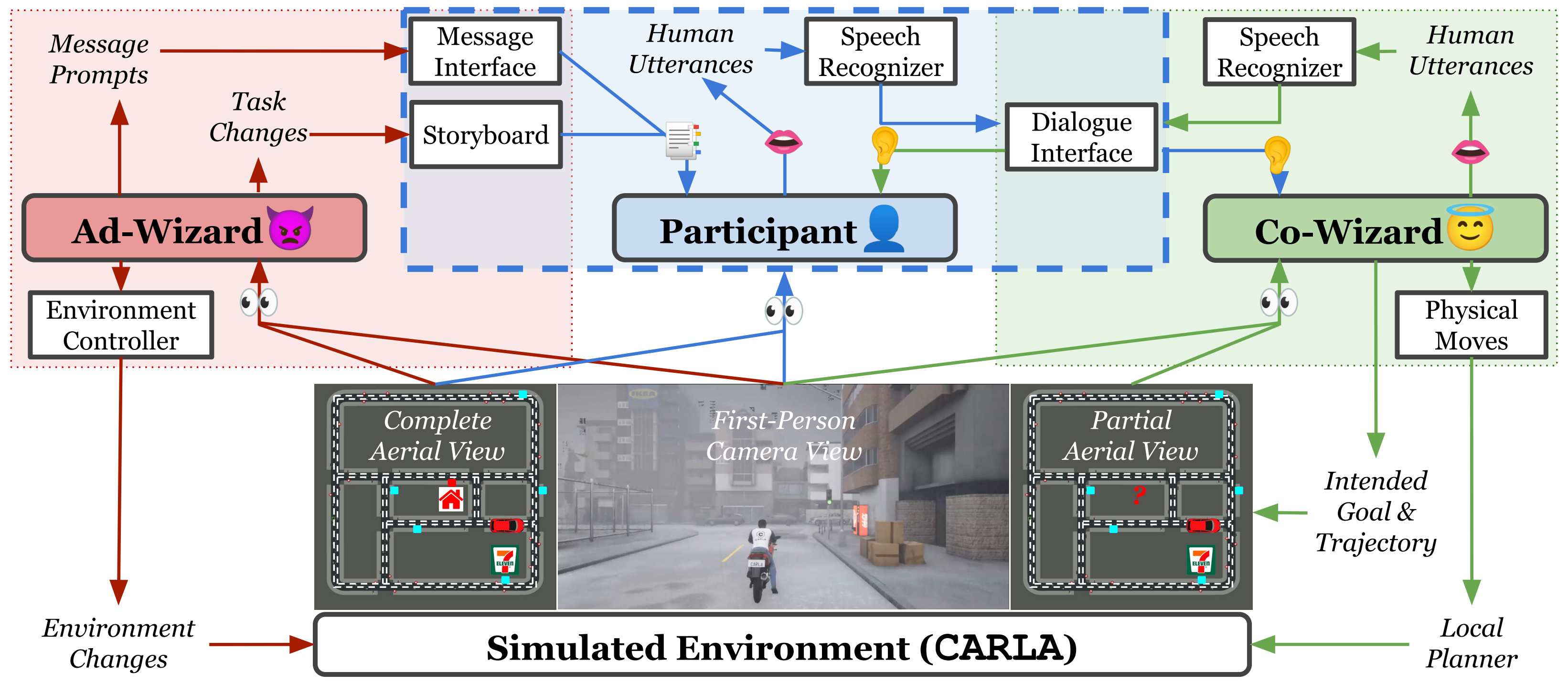




Simulated Platform

We adopt the **Dialogue On the ROad To Handle Irregular Events** (DOROTHIE), which was built upon CARLA [1] to study situated human-vehicle communication. We collect dialogue data in a **Wizard-of-Oz (WoZ)** setting.

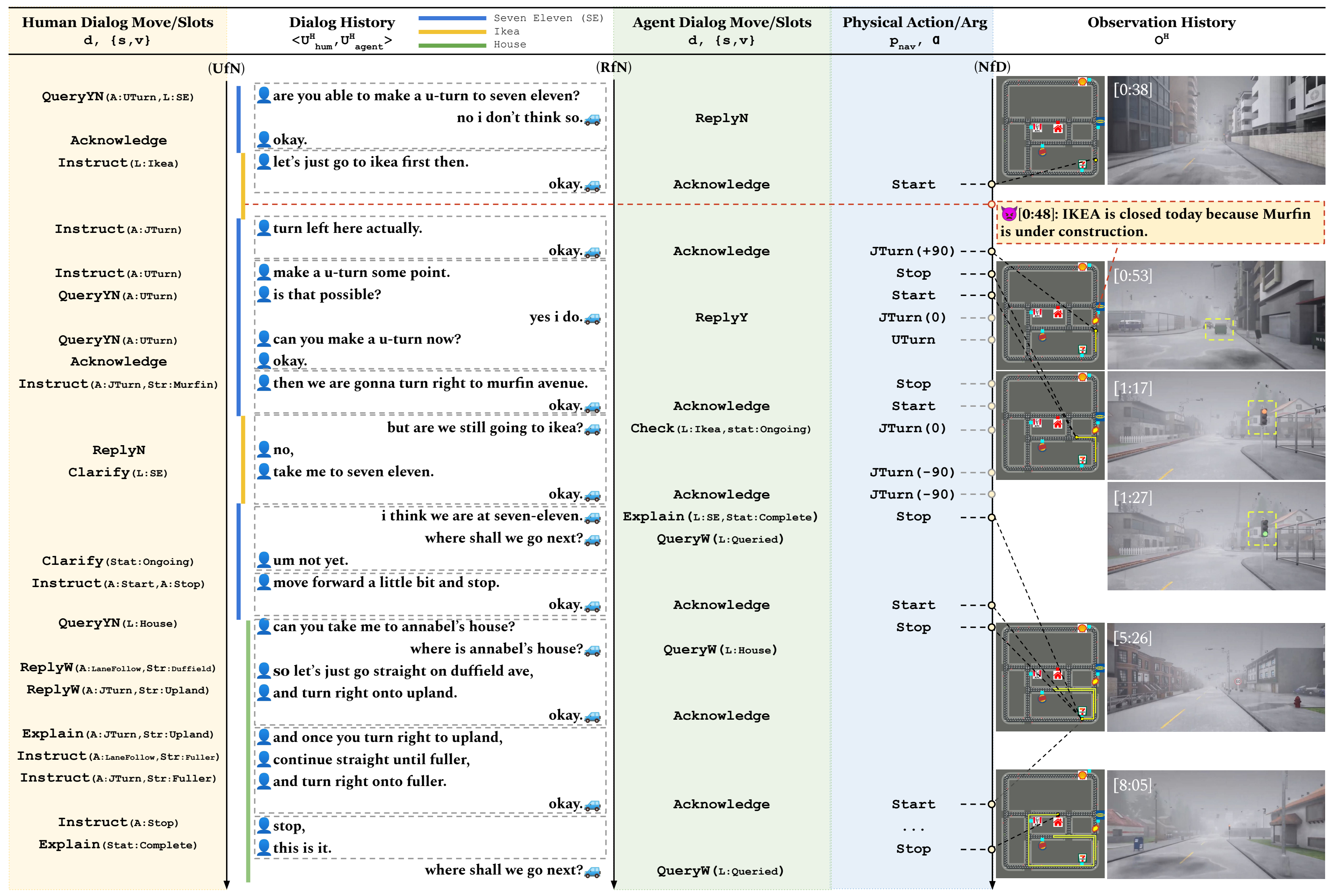
- The **participant** communicates with the vehicle to visit goal locations specified in a storyboard.
- The **Cooperative-Wizard** controls the agent's behaviors and carries language communication with the human participant to jointly achieve the goal.
- The **Adversarial-Wizard** controls the environment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



Task Definition and Data

We recruit 40 human subjects and collect Situated Dialogue Navigation (SDN) [2], a fine-grained navigation benchmark of 183 trials, consisting of over 8,000 utterances and 18.7 hours of control streams. We evaluate the agent's ability to generate dialogue and the next navigation actions.

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



👤 are you able to make a u-turn to seven eleven?
no i don't think so. 🚙
👤 okay.
👤 let's just go to ikea first then.
okay. 🚙
👤 turn left here actually.
okay. 🚙
👤 make a u-turn some point.
👤 is that possible?
yes i do. 🚙
👤 can you make a u-turn now?
👤 okay.
👤 then we are gonna turn right to murfin avenue.
okay. 🚙
but are we still going to ikea? 🚙
👤 no,
👤 take me to seven eleven.
okay. 🚙
i think we are at seven-eleven. 🚙
where shall we go next? 🚙
👤 um not yet.
👤 move forward a little bit and stop.
okay. 🚙
👤 can you take me to annabel's house?
where is annabel's house? 🚙
👤 so let's just go straight on duffield ave,
👤 and turn right onto upland.
okay. 🚙
👤 and once you turn right to upland,
👤 continue straight until fuller,
👤 and turn right onto fuller.
okay. 🚙
👤 stop,
👤 this is it.
where shall we go next? 🚙

👿[0:48]: IKEA is closed today because Murfin is under construction.

Besides these social experiences, we developed a data generation pipeline to obtain paired data of embodied perception and descriptions from the simulator. We replay the training sessions in the SDN benchmark to obtain the egocentric perception, record the environmental factors such as weather and nearby objects, and then fill these details into language descriptions using templates.

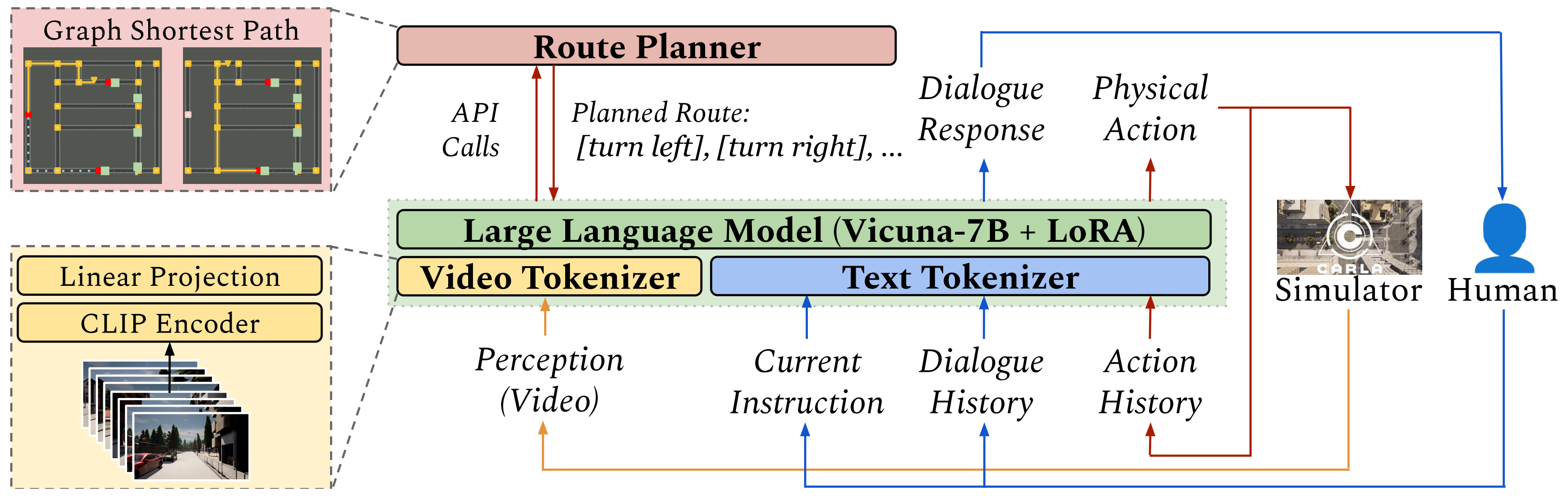
- **Distance to Road End**: We compute the distance to the road's end by subtracting the current waypoint's s value from the s value at the road's end.
- **Lane Information**: We note the lane number the car was in, counting from the left, and record whether the car could switch to the adjacent left or right lanes.
- **Object in Front**: We identify the object directly in front of the vehicle from the ground truth obtained from the simulation, and compute the distance to it.
- **Traffic Sign Visibility**: We record all visible traffic signs (e.g., traffic lights, stop signs, speed limit signs), along with the information they displayed (red/green for lights, posted speed limits), and their distances from the vehicle.
- **Weather Conditions**: We record the current weather conditions that could impact the vehicle's control.

DriVLMe Agent

The training process of DriVLMe consists of two stages:

- General video instruction tuning stage, focused on aligning the LLM and the video tokenizer using large-scale driving videos;
- Social and embodied instruction tuning stage, focused on training the LLM on the conversational data and episodes of embodied experiences in a simulator.

To enable symbolic planning for long-horizon goals, we introduce a route planner to incorporate the graph knowledge in the map into DriVLMe. The planner takes as input a given target landmark on the map and the current location of the agent and outputs a route from the agent to the target landmark following the shortest path.



Evaluation Results

Model	NfD		RfN			
	Act↑	Arg↑	Move↑	CIDEr↑	BERT↑	M↑
TOTO	41.2	36.0	40.9	-	-	-
GPT-4	53.0	44.2	11.0	0.06	0.48	0.09
GPT-4V	52.0	29.4	6.5	0.07	0.54	0.11
DriveVLM	**70.4**	**71.3**	61.4	0.43	**0.76**	**0.37**
DriVLMe (-social)	68.7	69.0	19.1	0.17	0.60	0.13
DriVLMe (-embodied)	68.4	67.7	**62.7**	**0.45**	**0.76**	**0.37**
DriVLMe (-domain)	62.4	70.7	60.9	0.35	0.75	0.18
DriVLMe (-video)	60.3	72.5	42.7	0.33	0.69	0.26
DriVLMe (-planner)	57.6	52.0	21.3	0.19	0.61	0.12

Model	Description			Justification			Full		
	C↑	B4↑	R↑	C↑	B4↑	R↑	C↑	B4↑	R↑
ADAPT	219.35	33.42	61.83	94.62	9.95	32.01	93.66	17.76	44.32
DriveGPT4	**254.62**	**35.99**	**63.97**	101.55	10.84	31.91	102.71	19.00	**45.10**
DriVLMe	227.05	33.39	61.02	**132.17**	**13.39**	**33.18**	**114.16**	**19.59**	44.83

Model	Speed					Turning Angle				
	E↓	A0.1↑	A0.5↑	A1↑	A5↑	E↓	A0.1↑	A0.5↑	A1↑	A5↑
ADAPT	3.02	9.56	24.77	37.07	90.39	11.98	27.93	66.83	75.13	89.45
DriveGPT4	**1.30**	**30.09**	**60.88**	**79.92**	98.44	**8.98**	59.23	**72.89**	**79.59**	**95.32**
DriVLMe	1.59	22.76	50.55	70.80	**99.20**	33.54	**61.38**	70.70	76.21	91.55

- Planner module contributes to response generation and the next action prediction.
- Social experiences significantly enhance response generation.
- Embodied experiences mainly aid the model in predicting actions unrelated to route planning, such as lane switching.

References

[1] *Dosovitskiy, Alexey, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun.* CARLA: An open urban driving simulator. In CoRL 2017.

[2] *Ma, Ziqiao, Ben VanDerPloeg, Cristian-Paul Bara, Yidong Huang, Eui-In Kim, Felix Gervits, Matthew Marge, and Joyce Chai.* DOROTHIE: Spoken Dialogue for Handling Unexpected Situations in Interactive Autonomous Driving Agents. In EMNLP Findings 2022.

Links to Demo