2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

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

Yidong Huang[1], Jacob Sansom[1], Ziqiao Ma[1], Felix Gervits[2], and Joyce Chai[1]
https://sled-group.github.io/driVLMe/

***Abstract*— Recent advancements in foundation models (FMs) have unlocked new prospects in autonomous driving, yet the experimental settings of these studies are preliminary, over-simplified, and fail to capture the complexity of real-world driving scenarios in human environments. It remains under-explored whether FM agents can handle long-horizon navigation tasks with free-from dialogue and deal with unexpected situations caused by environmental dynamics or task changes. To explore the capabilities and boundaries of FMs faced with the challenges above, we introduce DriVLMe, a video-language-model-based agent to facilitate natural and effective communication between humans and autonomous vehicles that perceive the environment and navigate. We develop DriVLMe from both embodied experiences in a simulated environment and social experiences from real human dialogue. While DriVLMe demonstrates competitive performance in both open-loop benchmarks and closed-loop human studies, we reveal several limitations and challenges, including unacceptable inference time, imbalanced training data, limited visual understanding, challenges with multi-turn interactions, simplified language generation from robotic experiences, and difficulties in handling on-the-fly unexpected situations like environmental dynamics and task changes. Nevertheless, DriVLMe offers a promising new direction for autonomous driving agents that need to navigate not just complex environments but also complex social interactions.**

I. Introduction

Autonomous driving (AD) has made remarkable progress in recent years, bringing us closer to a future where vehicles can function as our social robot partners that navigate roads safely and efficiently with minimal human intervention [1, 2]. As these AD agents start to enter our everyday lives, techniques to enable effective human-agent dialogue and collaboration become important. The ability to communicate with humans through natural language dialogue plays a crucial role in ensuring passenger safety, recovering from unexpected situations, gaining trustworthiness, and enhancing the overall driving experience [3, 4]. In traditional autonomous driving systems and in-vehicle dialogue systems, rule-based approaches [5–7] have been employed to interpret human instructions and generate appropriate responses. However, these systems often struggle to handle the complexity and variability of natural language, leading to limited functionality and sub-optimal performance. Recently, the paradigm has shifted to data-driven learning-based approaches [8–11], which offer language-based interpretability and promising results in short-horizon tasks.

Advances in foundation models (FMs) like Large Language Models (LLMs) have opened up new opportunities, as they demonstrate the ability to perform step-by-step reasoning [12], understand multimodal data [13], learn from embodied experiences [14, 15], and use external tools [16]. An increasing number of efforts [17–23] have demonstrated the potential of FMs in the field of autonomous driving. However, the experimental setups of these works are preliminary and simplified, compared to the driving scenarios in real human environments. One common limitation is the lack of an ability to handle long-horizon navigation tasks. Trained on simple action-level natural language instructions, these models perform well on short-horizon tasks like *turn* or *overtake* but fail to understand goal-level instructions that require route planning and map knowledge. Also, these systems only focus on following individual instructions in a single turn of interaction. Realistic interactions with human passengers often involve free-form dialogue, especially for collaboratively handling unexpected situations, e.g., those caused by sensor limitations, environmental dynamics, or task changes. Without modeling the interaction context, these models may fall short of understanding nuanced dialogue and providing appropriate responses in human-vehicle interactions.

To explore the capabilities and boundaries of FMs faced with the challenges above, we introduce DriVLMe, a novel video-language-model-based AD agent to facilitate natural and effective communication between humans and autonomous vehicles that perceive the environment and navigate. Motivated by Hu and Shu [24], our goal is to enhance a language model backend as world and agent models. We develop DriVLMe by learning from both *embodied experiences* in a simulated environment and *social experiences* from real human dialogue. Unlike previous works that only focus on open-loop benchmark evaluation using non-interactive datasets such as nuScenes [25] and BDD [26], we present both open-loop and closed-loop experiments in a simulated environment (CARLA [27]). For open-loop evaluations, we leverage the Situated Dialogue Navigation (SDN) [4] benchmark to assess DriVLMe's performance in generating dialogue responses and physical actions. Our experimental results have shown that DriVLMe significantly outperforms previous baselines on SDN by a large margin and competes with baselines trained with LLM-augmented data. We further conduct closed-loop pilot studies in the

*This work was supported by the Automotive Research Center (ARC) at the University of Michigan and NSF IIS1949634.
[1]University of Michigan, Ann Arbor, MI, 48109 USA. Contact: {owenhji, jhsansom, marstin, chaijy}@umich.edu
[2]Army Research Lab. Contact: felix.gervits.civ@army.mil

CARLA simulation environment in which DriVLMe performs a collaborative navigation task with human subjects. Our preliminary findings have demonstrated some promising abilities of DriVLMe in navigation and re-planning, and on the other hand also revealed several limitations including unacceptable inference time, imbalanced training data, and low image input resolution. We hope this paper offers a comprehensive perspective of the strengths and weaknesses of foundation models as AD agents while highlighting areas for future work.

II. Related Work

A. Foundation Models for Autonomous Driving

Recent research has explored the potential of LLMs in autonomous driving, e.g., by prompt engineering on off-the-shelf LLMs to obtain the driving decisions from textual descriptions of the surrounding environment [17, 23, 28], or by fine-tuning LLMs to predict the next action or plan future trajectories [29, 30]. To develop multimodal systems, both real and simulated driving videos have been utilized for instruction tuning [31]. For example, DriveGPT4 [20] and RAG-Driver [32] fine-tuned multimodal LLMs on real-world driving videos to predict future throttle and steering angles. DriveMLM [33] and LMDrive [18] adopted camera data and ego-vehicle states from the CARLA simulator. We refer to recent surveys and position papers for detailed reviews [34–37]. We note that the experimental setups in these efforts are preliminary and simplified compared to the real driving scenarios in human environments. First, these prior approaches were restricted to single human instructions (or even no language input), limiting performance on longer-horizon tasks with back-and-forth dialogue and higher-fidelity navigation goals. Furthermore, these prior models only focus on using LLMs to predict physical actions and give explanations, ignoring their potential to initiate dialogue and generate language responses from robotic experiences. Finally, none of these setups consider unexpected situations caused by sensor limitations, environmental dynamics, or plan changes, which are common in real-world environments.

B. Language-guided Autonomous Driving and Outdoor Vision-Language Navigation

Situated human-vehicle communication has been extensively studied in the form of spoken language, and this line of work dates back to early resources including several multilingual [38] and multimodal [39] speech corpora. Recently, vision-and-language navigation (VLN) tasks require an agent to navigate in a 3D environment based on natural-language instructions and egocentric camera observations, with some efforts in the outdoor scenarios [40, 41]. They consider the world as a discrete graph while agents navigate toward the goal by moving among nodes. Thanks to open-world autonomous driving simulators [27, 42, 43], recent work bridges the gap between discrete model prediction and continuous closed-loop control. Various language-guided autonomous driving experiments and datasets [4, 44, 45] have been developed based on these simulators.

C. Dialogue-guided Robotic Agents

Dialogue-guided agents for improving human-robot interaction have gained significant attention [46, 47]. Efforts in this field have ranged from enabling robots to adjust their plans in real-time based on human dialogue [48], to seeking additional hints [49, 50], or to ask for direct human collaboration [51] for task completion. For instance, InnerMonologue [52] investigates the use of LLMs for generating internal dialogue to assist in completing human-oriented tasks, while PromptCraft [53] explores precise prompt engineering to enhance the communication skills of robots. These developments underscore the pivotal role of foundation models as building blocks of agents to foster more effective human-robot collaboration.

III. Dorothie & Situated Dialogue Navigation

We set up our experiment in CARLA [27], a driving simulator for autonomous vehicles, and use the DOROTHIE framework [4] built upon it, which supports human-agent dialogue and various forms of unexpected situations. In this work, we adopt the problem definition and data from the SDN benchmark in [4].

A. Overview

The SDN benchmark is designed to assess the agent's capability in generating dialogue responses and physical navigation actions according to the perceptual and dialogue history. SDN is collected from human-human interactions in Wizard-of-Oz (WoZ) studies, consisting of over 8,000 utterances and 18.7 hours of control streams. In the WoZ study, a human participant engages with what they believe to be an autonomous driving agent to accomplish various navigation tasks. During the interaction, there is also an adversarial wizard who creates unexpected situations on the fly. This adversarial wizard changes environmental dynamics as well as current goals and plans by using language instructions and manipulating road conditions.

B. Problem Definitions

At time t, the agent is provided with a perceptual observation and a human language input, aggregated into the following model input:

- **Map knowledge.** A graph-structured topology M with a list of street names $\{\mathrm{str}_i\}$ and landmarks $\{\mathrm{lm}_i\}$.
- **Perceptual history.** A sequence of RGB images $V = \{V_0, V_1, \cdots, V_{t-1}\}$ captured by the first-person camera. The video sampling rate is 10Hz
- **Dialogue history.** The dialogue utterances from the human ($U_{t,\mathrm{HUM}}$) and the agent ($U_{t,\mathrm{BOT}}$).
- **Action history.** The action history includes a sequence of previous actions $A_t = \{a_0, a_1, \cdots, a_{t-1}\}$, where each action a_t is a tuple $\langle p, \alpha \rangle$ representing a physical action and its argument executed at time t. More details about physical action definitions are in Table I.

The goal of the agent is to navigate to a sequence of landmarks on the map following the dialogue instructions from the human partner. To guarantee coherence in future dialogues and unforeseen events, the tasks are defined

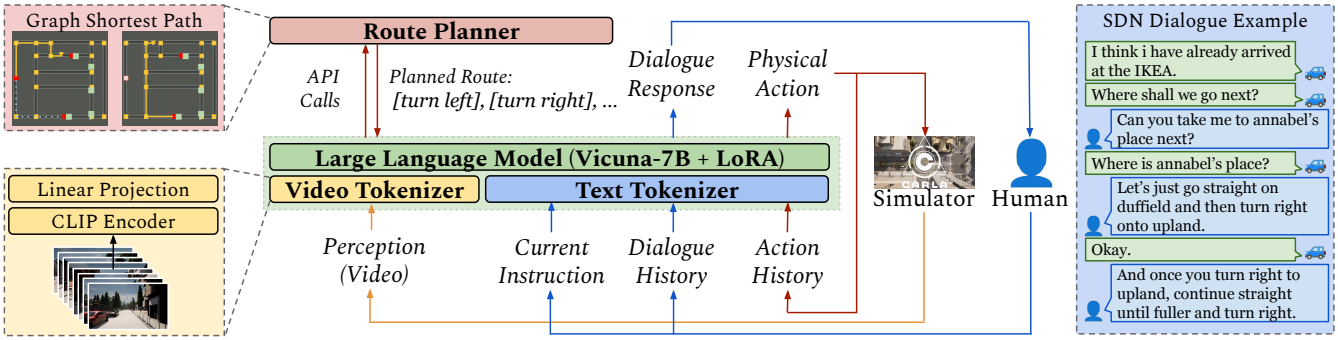

Fig. 1: Overview of the DriVLMe model architecture. DriVLMe is a multimodal Large Language Model that consists of (1) A video tokenizer that tokenizes the input visual history from the CARLA [27] simulator using a frozen CLIP encoder and a linear projection layer, (2) A route planner, a tool designed to assist the LLM in finding the shortest path from the agent's current location to another landmark specified by the LLM. (3) The base large language model, which receives input in the form of video representations, situated dialogue instructions, history of physical actions, and the output planned route from the route planner. It predicts dialogue responses to human inputs and physical actions that interact with the simulator.

TABLE I: The high-level action space in the SDN benchmark.

Physical Actions	Args	Descriptions
`LaneFollow`	-	Default behaviour, follow the current lane.
`LaneSwitch`	Direction	Switch to a neighboring lane.
`JTurn`	Direction	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.

in a teacher-forcing manner. This means that during data collection, the model is always presented with the actual action history A_t, rather than model-predicted actions during inference. The model is evaluated against the action and dialogue decisions of the human wizard. We particularly consider two sub-problems.

a) The Dialogue Response for Navigation (RfN) task.: The RfN task evaluates the agent's performance in generating an adequate response in driving-related communication. At time stamp τ, when the wizard makes an utterance, the agent is required to predict the dialogue response d. Instead of predicting only the dialogue move, we task the agent to generate the natural language.

b) The Navigation from Dialogue (NfD) task.: The NfD task evaluates the agent's performance in following human instructions from dialogue. At time stamp τ, when the wizard makes a decision on a physical action $\langle p, \alpha \rangle$, the agent is required to predict this physical action.

IV. Method

A. Model Architecture

Our DriVLMe agent is a large video-language model consisting of three parts: a video tokenizer, a route planning module, and an LLM backbone. The overview architecture of DriVLMe is visualized in Figure 1.

a) Video Tokenizer.: At time t, we can get a visual observation history $\{V_0, V_1, \cdots, V_{t-1}\}$. Given the long-range nature of the SDN benchmark, we assign a window size of $T_{\max} = 40$ with step $\Delta t = 2$ to sample the vision history and form a video $V \in \mathbb{R}^{T \times H \times W \times C}$, where H, W, and C are the height, width, and channel, respectively. For each video frame V_i, we adopt a pre-trained CLIP ViT-L/14 encoder [54] to extract the feature map $f \in \mathbb{R}^{T \times h \times w \times D}$, where $h = H/p$, $w = W/p$, p is the patch size of vision transformer, and D is the feature dimension of the CLIP encoder. We apply average-pooling to the feature map along the temporal dimension to get a representation $v_s \in \mathbb{R}^{(h \times w) \times D}$ and along the spatial dimensions to get a representation $v_t \in \mathbb{R}^{T \times D}$. By concatenating these two embeddings, we get the following video representation $v = \text{Concat}(v_t, v_s) \in \mathbb{R}^{(T+h \times w) \times D}$. We then use a linear projection layer g to project the embedding into the language decoder's embedding space with a dimension of K, resulting in the final embedding $g(v) = \mathbb{R}^{(T+h \times w) \times K}$.

b) LLM Backbone.: The LLM decoder is the core module that processes the input video and translates the dialogue instructions into lower-level decisions. Motivated by Video-ChatGPT [55], we adopt Vicuna-7B (v1.1) [56] as the LLM decoder. We also introduce a planning framework for environmental understanding with the detailed prompts shown in Figure 2.

c) Route Planning Module.: To enable symbolic planning for long-horizon goals, we introduce a route planner to incorporate the graph knowledge in the map M into DriVLMe. The planner takes as input a given target landmark on the map $\text{lm} \in \{\text{lm}_i\}$ and the current location of the agent l. It then outputs a route from the agent to the target landmark following the shortest path. To call the planner, the agent can simply output $\text{Plan}(\text{lm})$. The planner returns a list of turning directions, one per intersection in the route, expressed in natural language. The final output delivered to the DriVLMe agent is a list of directional action $\{p\} = [\text{dir}_1, \text{dir}_2, \cdots]$, where $\text{dir}_i \in \{\texttt{left}, \texttt{right}, \texttt{straight}, \texttt{uturn}\}$.

B. Instruction Tuning

Motivated by Hu and Shu [24], our goal is to enhance a language model's competence as a world model and agent model by learning from embodied experiences and social interactions. The training process of DriVLMe consists of two stages: (1) the general video instruction tuning stage, focused on aligning the LLM and the video tokenizer using

(**Video**)
(**System Message**): You are DriVLMe. You are responsible for safely piloting a car according to the instructions of a passenger. You must communicate with the passenger and make high-level decisions regarding the current navigational goals.
(**Prompt**): Describe what you see.
(**LLM**, Description): I can see a car in front of me. I can only switch left lane...
(**Dialogue & Action History**)
(**Route Planning Instruction**): You have a planning tool that you can plan your path to the destination. You can call it by `plan(destination)`, and it will return you a plan to get to your destination. If you don't have a destination in your mind, you can return `plan(None)`.
(**LLM**, Planning): `plan(ikea)`
(**Route Planner**): [left, straight, ...]
(**Prompt**): You can select a new navigational action and reply to the passenger.
(**LLM**, Action): `SwitchLane`
(**LLM**, Dialogue): "Ok, I will go to IKEA."

Fig. 2: **Example of system message and interaction between user and DriVLMe system.** The system message is an overview of the task the agent is required to accomplish. Given the video and the observation history, the agent is required to first describe the surrounding environment, then call the planner API to plan a route to the predicted goal, and make a decision at last. The output of the LLM is highlighted.

large-scale driving videos, and (2) the social and embodied instruction tuning stage, focused on training the LLM on the conversational data collected from real human-human dialogue and episodes of embodied experiences in a simulator.

1) Domain Video Instruction Tuning: Following the practice of Video-ChatGPT [55], we initialize the projection layer directly from LLaVA-7B (lightening v1.1) [57]. We adopt 50k video-text pairs from the BDD-X dataset [58] for the driving domain tuning. The pre-training images are collected from real driving videos and textual annotations of the environmental description and action explanations. We freeze the CLIP encoder and the LLM decoder, and train the projection layer only.

2) Social Instruction Tuning: At this stage, we used LoRA [59] to fine-tune the LLM in addition to the projector. We train the model on the whole training set of the SDN dataset, which has 13k video-dialogue pairs, including human-vehicle dialogues and long-term goals for planners. At each datapoint τ, the original SDN benchmark provides the dialogue d generated by human players, or physical action $\langle p, \alpha \rangle$, where p is an action (e.g., `Stop`) and α is an argument (e.g., `left`). We aim for the agent to learn how to plan in alignment with human intentions, which involves creating a sequence of primitive actions based on the goal and dialogue history, particularly when there is a change in the goal or plan. We manually annotate plan changes based on the car's trajectory and the current dialogue. While there could be several valid paths from the current location to the goal, we manually selected the routes that the vehicle took during the recording. These annotated plans serve as a part of the video-instruction data pairs for training, facilitating more effective learning of the planner as a tool.

3) Embodied Instruction Tuning: Besides the original dialogue data, 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.

4) Hyper-parameters.: The input resolution of the video is set as 224×224. We use a single linear layer for projection. For the pre-training stage of the model, we trained the model for 3 epochs with a learning rate of $2e^{-5}$ and a batch size of 4. We fine-tune the LLM with LoRA [59] and ZeRO [60]. The training epoch is 2 and the batch size is 1. For the LoRA configuration, we set rank to 128 and alpha to 256.

V. Open-Loop Evaluation

A. SDN Benchmark

For the open-loop evaluation, we tested the model on the test split of the SDN benchmark. The test set has two subsets, seen and unseen, where seen data points adopt either CARLA map Town01, Town03, or Town05 as the environment (which appeared in the training set). The unseen data points are from Town02, which is a relatively simple town map that was held out from training.

B. Evaluation Metrics

We evaluate our model on two tasks, RfN and NfD. The NfD task necessitates the agent's prediction of the physical action $\langle p, \alpha \rangle$, where p represents the chosen physical action and α is its argument. For evaluating both the physical action and its argument, we employ accuracy metrics. In the RfN task, the agent is required to predict the dialogue output d. The model is tasked with predicting the dialogue move m as defined in SDN. To evaluate the natural language dialogue output, we consider additional language generation metrics: CIDEr [61], BERTScore [62], and METEOR [63].

C. Baselines

a) Expert Baseline: We compared our model with TOTO [4], a baseline model implemented with an episodic transformer. Since the TOTO model does not have a text decoder and thus cannot generate dialogue, we only recorded the dialogue move prediction accuracy of TOTO.

b) Generalist Baselines: The GPT-4 [64] and GPT-4V [65] models are generalist LLMs we evaluated. Due to computational constraints, rather than test both models on the entirety of the SDN test set, we chose to randomly sample data points from four strata: seen RfN, unseen RfN, seen NfD, and unseen NfD. To evaluate each model on one of these strata, we randomly sampled 200 data points and fed them into a custom prompting infrastructure similar to the structure in Table 2. For the vision-enabled model (GPT-4V), we prepended an image V_{t-1} as the current visual input. To help the LLMs better understand the output format, we explain each option in the decision-making prompt.

TABLE II: Results of open-loop evaluation on the SDN test set. The seen sessions are from CARLA map Town01, Town03, and Town05, while unseen sessions are from CARLA map Town02. The NfD task measures the agent's ability to navigate according to human instruction and the RfN task measures the agent's ability to respond to humans in a situated dialogue, M stands for METEOR.

Model	NfD		RfN			
	Act↑	Arg↑	Move↑	CIDEr↑	BERT↑	M↑
Seen Environments						
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
DriVLMe	**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
Unseen Environment						
TOTO	45.8	41.1	31.0	-	-	-
GPT-4	67.5	61.3	14.5	0.05	0.47	0.08
GPT-4V	63.5	51.6	7.5	0.07	0.53	0.13
DriVLMe	70.8	**71.3**	**68.5**	**0.55**	**0.81**	**0.43**
DriVLMe (-social)	69.8	66.8	26.9	0.25	0.64	0.16
DriVLMe (-embodied)	**72.9**	68.0	66.7	0.52	0.79	0.42
DriVLMe (-domain)	65.9	70.8	65.3	0.48	0.78	0.38
DriVLMe (-video)	62.6	68.6	46.5	0.41	0.73	0.31
DriVLMe (-planner)	58.2	59.1	23.7	0.22	0.63	0.13

D. Main Results

As shown in Table II, our DriveVLMe model significantly outperformed the baseline models across most metrics, except for the physical action accuracy in the NfD task for the unseen map. This discrepancy may be attributed to the unfamiliarity with the unseen Town02, though it is topographically simpler. Overall, DriVLMe can predict more precise decisions and give better responses in situated dialogue compared to the baselines.

E. Ablation Studies

To assess the effectiveness of various data and components in developing DriVLMe, we conducted an ablation study. We evaluated the model performance by systematically removing specific training data and components to observe their impact on the model's ability to generate dialogue responses and predict physical actions.

- **Social Data (-social)**: We removed the human-vehicle dialogue data used for social instruction tuning.
- **Embodied Data (-embodied)**: We removed the simulated data used for embodied instruction tuning.
- **Domain Data (-domain)**: We removed the BDD-X data used for domain-general instruction tuning.
- **Video Input (-video)**: We removed the video processing component from DriVLMe and evaluated its performance without visual information.
- **Planner Module (-planner)**: We removed the planner module responsible for route planning in DriVLMe. This experiment aimed to assess the impact of proactive route planning on the model's navigation capabilities.

As shown in Table II, removing the video input and the planner module both decrease the performance of the model on the RfN tasks on all metrics, indicating the contribution of both modules on response generation. A similar decrease in NfD performance is observed, while the impact of removing the planner is significant, suggesting that the route planner module greatly contributes to the next action prediction. Data ablation studies show that social experiences significantly enhance response generation. We observed that embodied experiences mainly aid the model in predicting actions unrelated to route planning, such as lane switching. Consequently, this was less beneficial in the unseen Town02, where lane switching is not necessary.

VI. Closed-loop Evaluation

For the closed-loop evaluation, we developed a human-in-the-loop simulation protocol in CARLA based on the simulator developed in DOROTHIE for human studies.

A. Experimental Design

We designed our closed-loop experiment to assess the adaptability and robustness of our autonomous driving system under various dynamic scenarios. The experiment was conducted in Town01 and Town02, including both seen and unseen maps. A human subject instructed the DriVLMe agent to navigate to a preset goal by giving natural language instructions following the storyboard, and the agent attempted to follow these instructions, autonomously navigate in the environment, and communicate with the human subject. To comprehensively evaluate the system's performance, we test the model with different settings as specific in the storyboards below:

- **Long-horizon v.s. Short-horizon Instructions**: Users instruct the agent with either long-horizon instructions, involving higher-level navigational goals (e.g., "go to the KFC"), or short-horizon instructions (e.g., "turn right at the next intersection") asking for immediate maneuvers.
- **Weather Change**: A sudden weather change (e.g., rain) is triggered during driving.
- **Goal Change**: The human user asks for a change of goal to let the agent replan the route. The human user first instructs the agent to navigate to an initial goal and then updates it.
- **Obstacle Addition**: An obstacle is placed in front of the agent to force a stop or lane change.

B. Connecting DriVLMe to Simulation

Throughout 20 pilot studies with real human subjects, agents' interactions with the simulator formed a closed-loop control mechanism. We used a local motion planner to translate the physical actions back into throttle and steering control. Due to the LLM inference rate, we limited the LLM to interact with the environment at a frequency of 2 Hz, and provided the model with the whole interaction history H_t to prompt the model. For the evaluation, we used whether the final goal was achieved as the metric and recorded the failure cases for analysis.

Fig. 3: Examples of closed-loop evaluation of DriVLMe in CARLA, following action-level natural language instructions.

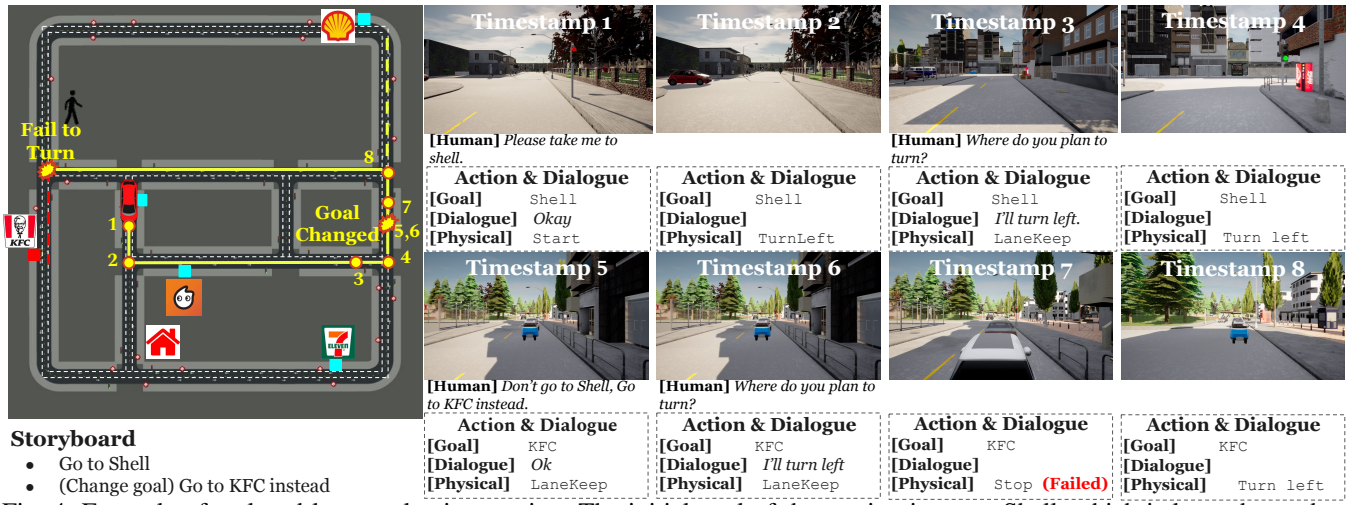

Fig. 4: Example of a closed-loop evaluation session: The initial goal of the session is set to Shell, which is later changed to KFC during the course of the evaluation. The yellow solid line represents the path taken by the agent and the yellow dotted line represents the route planned by the planner. We took eight checkpoints in the whole evaluation session and recorded the input dialogue, goal prediction, dialogue response and the physical action taken for each checkpoint.

C. Main Results

The outcomes of our experimental investigations show the promise of the DriVLMe model in autonomous driving dialogue tasks (with 6 successful sessions out of 20 tests), whereas the failure cases highlight areas for future work. As can be seen in Figure 3, we find that the DriVLMe model is capable of following simple human instructions and performing the physical actions as requested, in line with previous studies on foundation model agents for autonomous driving. Surprisingly, we find that DriVLMe can effectively call the route planner API for reliable graph planning and re-planning, demonstrating LLMs' tool use capabilities. The model is also robust under weather changes during the session. Still, these successful sessions are limited to cases when there is one single long-horizon goal or only one change of goal. We observe challenges with multi-turn interactions with multiple short-horizon instructions. DriVLMe also faces difficulties in handling unexpected situations and changes to environmental dynamics. Lastly, the simplified language generation from robotic experiences has triggered concerns about trustworthiness as raised by human subjects. Figure 4 shows an example of our session with a goal change instruction. We find that the agent can react to goal changes and plan turns according to the plan given by the route planner tool. However, we encountered two failure cases during the experiment. First, the agent failed to stop when the car in front suddenly stopped (timestamp 7). Second, the agent failed to predict a turn at the last intersection, causing the agent to stall at the intersection (as marked on the map). We present the video demonstration for additional details and discuss the limitations of foundation model agents in the following section.

VII. Limitations and Future Work

Our pilot studies revealed several failure cases and technical challenges for LLM-based AD agents, outlined as follows.

a) Imbalanced Embodied Experiences: The imbalanced training data in autonomous driving tasks, where the majority of data points are routine actions like lane following, leads to model biases and poor performance on less frequent actions such as *stop*. Addressing this issue requires introducing robust data augmentation in embodied experiences, sampling strategies, or domain-specific knowledge injected throught the training process.

b) Limited World Modeling and Visual Understanding: Low image resolution and the absence of OCR capabilities hinder the visual encoder's ability to capture critical world states, leading to misinterpretations of traffic lights and signs. Enhancing image resolution, integrating OCR, and incorporating complementary sensors into LLMs could improve perception.

c) Unexpected Situations and World Dynamics: Our experiments also shows that the agent struggles with out-of-distribution corner cases while these cases are common in real-world driving setting. Future work could involve learning from in-the-wild driving video/data or enabling LLMs to seek human help in unforeseen circumstances.

d) Language Generation from Embodied Experiences: The model's language output is often oversimplified, primarily consisting of straightforward responses to human instructions or simplistic yes/no replies. It also fails to initiate dialogues or request additional guidance from human. Future work should focus on enhancing the model's conversational initiative, enabling self-motivated dialogue.

e) Multi-turn Interactions and Instruction Following: The agent occasionally loses track of long-horizon instructions during extended dialogues, leading to incorrect goal predictions. This issue underscores the critical importance of memory retention and context awareness in maintaining an agent model, particularly in situations where extensive dialogue exchange happens. This challenge could be addressed through implementation of memory-based mechanisms within LLM architectures or adding some memory modules in the autonomous driving agent framework.

f) Limited Theory of Mind and Trust-worthiness: Another critical limitation observed in our study is the absence of a situated Theory of Mind (ToM) [66] in the autonomous agent. The agent sometimes misinterprets low-level instructions, mistaking them as cues to abandon long-horizon goals. Developing a nuanced understanding of the instructor's intentions is vital for building trust and improving interaction modeling.

VIII. Conclusion

In this work, we presented DriVLMe, an LLM-based autonomous driving agent that leverages both embodied experiences in a simulated environment and social experiences in real human dialogue. The egocentric perception and conversational interaction empower DriVLMe to engage in meaningful dialogues with human passengers while navigating complex driving environments. Through empirical evaluations, we demonstrated the effectiveness and versatility of DriVLMe in autonomous driving dialogue tasks, showcasing significant improvements in both physical action prediction and dialogue response generation metrics. Our findings demonstrate the potential of DriVLMe in enabling human-agent communication and autonomous driving, while also highlighting limitations and areas of future work.

References

[1] W. Schwarting, A. Pierson, J. Alonso-Mora, S. Karaman, and D. Rus, "Social behavior for autonomous vehicles," *Proceedings of the National Academy of Sciences*, vol. 116, no. 50, pp. 24 972–24 978, 2019.

[2] W. Wang, L. Wang, C. Zhang, C. Liu, L. Sun, *et al.*, "Social interactions for autonomous driving: A review and perspectives," *Foundations and Trends® in Robotics*, vol. 10, no. 3-4, pp. 198–376, 2022.

[3] F. Weng, P. Angkititrakul, E. E. Shriberg, L. Heck, S. Peters, and J. H. Hansen, "Conversational in-vehicle dialog systems: The past, present, and future," *IEEE Signal Processing Magazine*, vol. 33, no. 6, pp. 49–60, 2016.

[4] Z. Ma *et al.*, "DOROTHIE: Spoken dialogue for handling unexpected situations in interactive autonomous driving agents," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates, 2022, pp. 4800–4822.

[5] B. Pellom *et al.*, "University of colorado dialogue systems for travel and navigation," in *Proceedings of the first international conference on Human language technology research*, 2001.

[6] J. Baca, F. Zheng, H. Gao, and J. Picone, "Dialog systems for automotive environments.," in *INTERSPEECH*, 2003, pp. 1929–1932.

[7] W. Schwarting, J. Alonso-Mora, and D. Rus, "Planning and decision-making for autonomous vehicles," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 1, pp. 187–210, 2018.

[8] A. Kendall *et al.*, "Learning to drive in a day," in *2019 international conference on robotics and automation (ICRA)*, IEEE, 2019, pp. 8248–8254.

[9] Y. Hu *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 853–17 862.

[10] B. Jin *et al.*, "Adapt: Action-aware driving caption transformer," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2023, pp. 7554–7561.

[11] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[12] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[13] Y. Zhang, Z. Ma, X. Gao, S. Shakiah, Q. Gao, and J. Chai, "Groundhog: Grounding large language models to holistic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.

[14] Y. Mu *et al.*, "Embodiedgpt: Vision-language pre-training via embodied chain of thought," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[15] J. Xiang *et al.*, "Language models meet world models: Embodied experiences enhance language models," *Advances in neural information processing systems*, vol. 36, 2023.

[16] T. Schick *et al.*, "Toolformer: Language models can teach themselves to use tools," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[17] L. Wen *et al.*, "Dilu: A knowledge-driven approach to autonomous driving with large language models," in *The Twelfth International Conference on Learning Representations*.

[18] H. Shao *et al.*, "Lmdrive: Closed-loop end-to-end driving with large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 120–15 130.

[19] X. Tian *et al.*, "Drivevlm: The convergence of autonomous driving and large vision-language models," *arXiv preprint arXiv:2402.12289*, 2024.

[20] Z. Xu *et al.*, "Drivegpt4: Interpretable end-to-end autonomous driving via large language model," *arXiv preprint arXiv:2310.01412*, 2023.

[21] Y. Jin *et al.*, "Surrealdriver: Designing generative driver agent simulation framework in urban contexts based on large language model," *arXiv preprint arXiv:2309.13193*, 2023.

[22] Y. Ma, Y. Cao, J. Sun, M. Pavone, and C. Xiao, "Dolphins: Multimodal language model for driving," *arXiv preprint arXiv:2312.00438*, 2023.

[23] H. Sha *et al.*, "Languagempc: Large language models as decision makers for autonomous driving," *arXiv preprint arXiv:2310.03026*, 2023.

[24] Z. Hu and T. Shu, "Language models, agent models, and world models: The law for machine reasoning and planning," *arXiv preprint arXiv:2312.05230*, 2023.

[25] H. Caesar *et al.*, "Nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[26] F. Yu *et al.*, "Bdd100k: A diverse driving dataset for heterogeneous multitask learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2636–2645.

[27] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*, PMLR, 2017, pp. 1–16.

[28] D. Shah, B. Osiński, S. Levine, *et al.*, "Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action," in *Conference on robot learning*, PMLR, 2023, pp. 492–504.

[29] L. Chen *et al.*, "Driving with llms: Fusing object-level vector modality for explainable autonomous driving," *arXiv preprint arXiv:2310.01957*, 2023.

[30] J. Mao, Y. Qian, J. Ye, H. Zhao, and Y. Wang, "Gpt-driver: Learning to drive with gpt," in *NeurIPS 2023 Foundation Models for Decision Making Workshop*, 2023.

[31] C. Sima *et al.*, "Drivelm: Driving with graph visual question answering," in *First Vision and Language for Autonomous Driving and Robotics Workshop*.

[32] J. Yuan *et al.*, "Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model," *arXiv preprint arXiv:2402.10828*, 2024.

[33] W. Wang *et al.*, "Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving," *arXiv preprint arXiv:2312.09245*, 2023.

[34] X. Li *et al.*, "Towards knowledge-driven autonomous driving," *arXiv preprint arXiv:2312.04316*, 2023.

[35] C. Cui *et al.*, "A survey on multimodal large language models for autonomous driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 958–979.

[36] H. Gao, Y. Li, K. Long, M. Yang, and Y. Shen, "A survey for foundation models in autonomous driving," *arXiv preprint arXiv:2402.01105*, 2024.

[37] X. Yan *et al.*, "Forging vision foundation models for autonomous driving: Challenges, methodologies, and opportunities," *arXiv preprint arXiv:2401.08045*, 2024.

[38] H. van den Heuvel, J. Boudy, R. Comeyne, S. Euler, A. Moreno, and G. Richard, "The speechdat-car multilingual speech databases for in-car applications: Some first validation results," in *EUROSPEECH*, 1999.

[39] B. Lee *et al.*, "Avicar: Audio-visual speech corpus in a car environment," in *Eighth International Conference on Spoken Language Processing*, 2004.

[40] A. B. Vasudevan, D. Dai, and L. Van Gool, "Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 246–266, 2021.

[41] J. Li, A. Padmakumar, G. Sukhatme, and M. Bansal, "Vln-video: Utilizing driving videos for outdoor vision-and-language navigation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.

[42] M. Zhou *et al.*, "Smarts: An open-source scalable multi-agent rl training school for autonomous driving," in *Conference on robot learning*, PMLR, 2021, pp. 264–285.

[43] E. Vinitsky, N. Lichtlé, X. Yang, B. Amos, and J. Foerster, "Nocturne: A scalable driving benchmark for bringing multi-agent learning one step closer to the real world," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3962–3974, 2022.

[44] N. Sriram, T. Maniar, J. Kalyanasundaram, V. Gandhi, B. Bhowmick, and K. M. Krishna, "Talk to the vehicle: Language conditioned autonomous navigation of self driving cars," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2019, pp. 5284–5290.

[45] J. Roh, C. Paxton, A. Pronobis, A. Farhadi, and D. Fox, "Conditional driving from natural language instructions," in *Proceedings of the Conference on Robot Learning*, 2020, pp. 540–551.

[46] M. Marge *et al.*, "Spoken language interaction with robots: Recommendations for future research," *Computer Speech & Language*, vol. 71, p. 101 255, 2022.

[47] T. Minato, R. Higashinaka, K. Sakai, T. Funayama, H. Nishizaki, and T. Nagai, "Design of a competition specifically for spoken dialogue with a humanoid robot," *Advanced Robotics*, vol. 37, no. 21, pp. 1349–1363, 2023.

[48] P. Sharma *et al.*, "Correcting robot plans with natural language feedback," *arXiv preprint arXiv:2204.05186*, 2022.

[49] J. Thomason, M. Murray, M. Cakmak, and L. Zettlemoyer, "Vision-and-dialog navigation," in *Conference on Robot Learning*, PMLR, 2020, pp. 394–406.

[50] A. Padmakumar *et al.*, "Teach: Task-driven embodied agents that chat," in *AAAI*, 2022.

[51] K. X. Nguyen, Y. Bisk, and H. D. Iii, "A framework for learning to request rich and contextually useful information from humans," in *International Conference on Machine Learning*, PMLR, 2022, pp. 16 553–16 568.

[52] W. Huang *et al.*, "Innermonologue: Embodied reasoning through planning with language models," CoRL 2022 (to appear), 2022.

[53] A. Z. Ren *et al.*, "Robots that ask for help: Uncertainty alignment for large language model planners," in *Conference on Robot Learning*, PMLR, 2023, pp. 661–682.

[54] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[55] M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024.

[56] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[57] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," vol. 36, 2023.

[58] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[59] E. J. Hu *et al.*, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations*.

[60] S. Rajbhandari, J. Rasley, O. Ruwase, and Y. He, "Zero: Memory optimizations toward training trillion parameter models," in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, IEEE, 2020, pp. 1–16.

[61] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[62] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2019.

[63] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[64] J. Achiam *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[65] OpenAI, *Gpt-4v(ision) system card*, Sep. 2023.

[66] Z. Ma, J. Sansom, R. Peng, and J. Chai, "Towards a holistic landscape of situated theory of mind in large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 1011–1031.