doScenes: An Autonomous Driving Dataset with Natural Language Instruction for Human Interaction and Vision-Language Navigation

Parthib Roy, Srinivasa Perisetla, Shashank Shriram, Harsha Krishnaswamy, Aryan Keskar, Ross Greer

Abstract—Human-interactive robotic systems, particularly autonomous vehicles (AVs), must effectively integrate human instructions into their motion planning. This paper introduces doScenes, a novel dataset designed to facilitate research on human-vehicle instruction interactions, focusing on short-term directives that directly influence vehicle motion. By annotating multimodal sensor data with natural language instructions and referentiality tags, doScenes bridges the gap between instruction and driving response, enabling context-aware and adaptive planning. Unlike existing datasets that focus on ranking or scenelevel reasoning, doScenes emphasizes actionable directives tied to static and dynamic scene objects. This framework addresses limitations in prior research, such as reliance on simulated data or predefined action sets, by supporting nuanced and flexible responses in real-world scenarios. This work lays the foundation for developing learning strategies that seamlessly integrate human instructions into autonomous systems, advancing safe and effective human-vehicle collaboration. We make our data publicly available at https://www.github.com/rossgreer/doScenes

Index Terms—safe autonomous driving, human-robot interaction, vision language action models, motion planning

I. INTRODUCTION

THERE is a growing need for robotic systems, especially autonomous vehicles, to be human-interactive. In this research, we particularly focus on *human-vehicle instruction interactions*, where a human agent communicates a directive to a vehicle that should influence the vehicle's motion plan. While many of the principles discussed in this research extend more generally to human-robot instruction interactions; we focus on autonomous vehicles as a special case of robot whose motion plans exists in a particular scale of time and velocity, necessitating but also benefiting from domain-specific characterizations of instructions.

Existing interactions of humans and vehicles can be characterized by a set of attributes such as source position [1], modality, referentiality, and temporality. Example options within these attributes are summarized in Table I.

Instructions may be described by combinations of these attributes, and options within an attribute are not always mutually exclusive and may be integrated in various combinations. For example:

- A passenger may point to a curb cut and ask to be dropped off there, using verbal and gesture-based interaction from inside the vehicle, and providing a short-term instruction which refers to a static object in the scene.
- A firefighter may ask a vehicle to move out of the way, using verbal instruction from outside the vehicle,



1

"Turn left, then slow down and turn slightly right to avoid road obstruction." Static-Referential

Fig. 1. Typical nuScenes data includes 3D bounding box annotations, LiDAR point clouds, and driving area map feature layers. In the doScenes dataset, we augment each clip of temporal data with an instruction and a tag to indicate the instruction's referentiality.

and providing a short-term instruction which refers to dynamic scene objects.

• A police officer may use their whistle and hand-gestures to get a driver's attention and wave their vehicle through while directing traffic, using a combination of pseudoverbal and gesture-based interaction from outside the vehicle, and providing a short-term instruction which does not refer to additional objects.

In these examples, and in this research, we focus on shortterm interactions, which anecdotally apply on the order of less than 10 seconds of motion. More specifically, these types of instructions contain all relevant information within a viewable proximity (e.g. no relevant landmarks or information beyond the horizon of the driver's egocentric view). The time itself is not a strictly-defined boundary. For example, in the nuScenes dataset, samples have 12 seconds of motion; sometimes, these 12 seconds stay within a visible horizon from the temporal origin, and other times, the vehicle effectively moves to an entirely new scene within 12 seconds¹. Towards the development of new learning strategies in autonomous perception and planning, we introduce the doScenes dataset, a novel dataset which pairs sensor feeds, vehicle trajectories, and map information with human-interactive instructions and referentiality tags. doScenes draws its name from Judea Pearl's docalculus [2], developed to identify causal effects; accordingly, the human instructions provided in this dataset are intended

¹This point is discussed further in the section of the paper *Considerations for Application and Evaluation.*

Corresponding author R. Greer is with the Department of Computer Science and Engineering, University of California, Merced. e-mail: ross-greer@ucmerced.edu. All authors are members of the Machine Intelligence, Interaction, and Imagination Lab (Mi³) at the University of California, Merced.

 TABLE I

 Attributes of Human-Vehicle Instruction Interactions

Source Position	Inside Vehicle, Outside Vehicle
Modality	Voice, Gesture
Referentiality	None, Static Objects, Dynamic Objects
Temporality	Short-Term, Medium-Term, Long-Term

to emulate commands that would cause the resulting sequence of actions taken by the driver.

II. RELATED RESEARCH

A. Datasets for Human-Robot Instruction

Before discussing datasets specific to autonomous driving, we first present related research in the more general field of Human-Robot Interaction (HRI), and specifically instructionstyle interaction for real-world scenarios. One such dataset is NatSGD [3], a multimodal dataset designed to emulate natural human communications through speech and gestures. NatSGD is primarily designed to enable robots to understand and execute real-world tasks in a natural manner, including those requiring nuanced household robotics actions like cooking and cleaning. It stands out as one of the first datasets to encompass speech, gestures, and demonstration trajectories. In the NatSGD framework, robot behaviors were developed by creating a photorealistic simulated environment using Unity3D in conjunction with a customized Robot Operating System (ROS) plugin. Additionally, real-time inverse kinematics for the robot's head movement and arms were implemented using BioIK. During tasks, the robot maintained eye contact with the target and returned its gaze to the participant when ready for the next interaction.

Another high-volume and diverse dataset is BridgeData V2 [4]. BridgeData V2 distinguishes itself by offering coverage across numerous tasks and domains in robotic learning research, including support for task conditioning through goal images or natural language instructions. The dataset includes over 60,000 trajectories (50,365 expert demonstrations and 9,731 from a randomized scripted policy) collected across 24 environments. Data collection used an accessible, lowcost robot platform, making BridgeData V2 appealing for academic research. The robot setup consisted of a WidowX 250 robot arm (fixed-based) and numerous cameras, including one RGBD camera for sensing, two RGB cameras with randomized poses during data collection, and one RGB camera attached to the robot's wrist. In addition, the robot was controlled via a VR controller. Data collection occurred in indoor environments, with the majority being in toy kitchens. However, despite its scale and variety of tasks, the dataset is limited to low-precision and non-real-time activities, quite different from the requirements of autonomous driving.

The HandMeThat [5] benchmark assesses instruction understanding and task execution within physical and social contexts, emphasizing situations and instructions with ambiguity. Each episode of the text-based dataset contains a sequence of steps taken by the human, followed by an instruction. The authors propose two stages for modeling robot responses. In the first stage, a robot agent observes a human agent and its actions and attempts to infer their end goal; in the second stage, the human provides a language-based instruction to the robot, and the robot acts within the environment to complete its tasks. The robot agent needs to consider both the human's historical actions as well as the subgoal specific to human utterance. It is important to note that the benchmark lies within its operation of a text-only environment, strongly limiting its scope for vision-based environments, and does not address non-verbal communication or dynamic interactions.

B. nuScenes and Natural Language in Autonomous Driving Datasets

nuScenes [6] is a multimodal dataset designed to fill the gap of capturing diverse real-world conditions necessary for building robust autonomous driving perception systems. At the time of its release, nuScenes was the largest AV dataset to feature a complete 360° field of view (FOV) AV sensor suite, including 6 cameras, 5 radars, and 1 lidar, and is also the first to include radar data using an AV approved for public roads. Additionally, nuScenes was the first multimodal dataset to capture nighttime and rainy condition data. The dataset includes 1,000 manually selected scenes from two highly challenging and dense traffic environments: Boston (Seaport and South Boston) and Singapore (One North, Holland Village, and Queenstown). Each scene is annotated at 2 Hz, resulting in 1.4 million 3D bounding boxes for 23 object classes. The AV utilized during data collection were two Renault Zoe supermini electric cars, equipped with front and side cameras with a 70° FOV, offset by 55°, and a rear camera with a 110° FOV.

Though novel in its instruction and interactivity basis, doScenes is not the first dataset which features nuScenes annotations extended using natural language. nuScenes-QA [7] combined nuScenes' 3D detection annotations with question templates, automatically generating 460K question-answer pairs based on scene graphs. nuScenes-MQA (Markup Question Answering) [8] introduced questions and answers enclosed within markups of particular objects within the visual scene.

Beyond re-annotations of nuScenes, other datasets have been developed to integrate natural language information into the driving environment. The Rank2Tell dataset [9] advances autonomous driving with multimodal data annotated with visual elements in a traffic scene ranked by relevance to safety, traffic rule compliance, and the dynamic context of the situation. By emphasizing contextual prioritization, Rank2Tell provides a benchmark for evaluating how well autonomous vehicle systems align with human judgment. While Rank2Tell makes significant contributions to ranking-based reasoning and highlights the importance of competing visual elements, it is limited to scene-level understanding and is only based on certain specific key-frames of the traffic scene video. It lacks actionable instructions for motion planning, restricting its applicability to real-world driving scenarios where vehicles must respond to specific directives. doScenes addresses this gap by bridging the divide between multimodal reasoning and actionable instructions. Annotations can be directly tied to objects of importance in traffic scenes, focusing on the execution of human and natural-language commands. By complementing Rank2Tell's emphasis on importance ranking with actionable directives, doScenes offers a novel framework for training and evaluating autonomous systems in real-world human-vehicle interactions.

The GPT-Driver framework [10] transforms autonomous driving motion planning into a language modeling task using OpenAI's GPT-3.5 model. It converts inputs like sensor data and vehicle states into a unified language representation, generating driving trajectories as language tokens with natural language explanations. This tokenized-driving approach enhances interpretability and generalization, allowing for greater transparency in decision-making. Tested on the nuScenes dataset, GPT-Driver achieved state-of-the-art trajectory prediction accuracy on the nuScenes dataset with a centimeterlevel L2 error and competitive collision rates. However, it is evaluated in an open loop form on 3-second intervals, and does not make use of important features found in doScenes, such as egocentric views or human instruction beyond generic highlevel objectives (e.g. "right"), limiting its ability to generalize to novel scenarios or instructions.

With similar data to doScenes, the DriveMLM and LMDrive frameworks [11], [12] leverage large language models (LLMs) for autonomous driving by integrating multimodal data from CARLA simulations, including images, LiDAR data, traffic rules, and user commands, to align with human instructions. The primary aim is to enable the LLM to predict actionable steps for the ego vehicle to execute given specific human commands, a similar objective to doScenes. DriveMLM is supported by a robust dataset of 280 hours of annotated CARLA simulation data, including decision states and corresponding explanations. LMDrive enables vehicles to perform step-by-step navigation tasks, supported by a 64K-clip dataset encompassing diverse scenarios and complex instructions. LMDrive segments training data into clips, with each clip corresponding to one navigation instruction from a pre-defined set of 56 instructions. Notably, they augment their dataset by using ChatGPT to generate 8 semantically-equivalent variants of each instruction.

However, the reliance of DriveMLM and LMDrive on simulated data restricts their applicability to real-world scenarios. Additionally, the frameworks prioritize immediate, one-step decisions over multi-step planning or adaptive driving styles necessary for handling complex environments, reflected in the limited output of DriveMLM within a set of instructions such as keep, accelerate, left change, right change, which may fail to fully capture the nuances and complexities of diverse driving scenarios and constrain the system's flexibility in responding to unique situations requiring dynamic or nuanced decision-making. doScenes does not limit instructions to a fixed set; annotators can freely describe each scene's instructions, and because there are multiple ways to give an instruction with the same effect, some choose to make references to scene objects while others do not. doScenes brings the concepts of DriveMLM and LMDrive to real-world data and provides instructions which serve decisions beyond one-step decisions, further augmented with tagged information on whether instructions rely on dynamic or static objects

within the driving scenes. We note that, at the time of writing, only LMDrive has made their data publicly available.

DriveGPT4 [13] is an interpretable end-to-end autonomous driving system using LLMs for multimodal reasoning and control. It integrates video inputs and textual queries to predict low-level vehicle control signals and provides natural language explanations for its actions. Motivating our research, the authors note the scarcity of publicly available datasets suitable for their task, and train on an enhanced BDD-X dataset with ChatGPT-generated Q&A pairs for questions equivalent to *what* is the vehicle doing, *why* is the vehicle doing it, and *what will* the vehicle do next. DriveGPT4 excels in action description, justification, and control signal prediction, offering actionable decisions and user-friendly explanations. However, the dataset is again without human instruction, only explanation of the current state.

The DRAMA dataset [14] advances situational awareness in autonomous driving by addressing two key challenges: identifying risks and explaining them. DRAMA (Driving Risk Assessment Mechanism with A captioning module) distinguishes itself by not only pinpointing risks in driving scenes but also describing them in natural language. It contains 17,785 real-world driving scenarios from urban roads in Tokyo, Japan, with detailed annotations on risks, critical objects, and interactions from the driver's perspective. By combining visual reasoning with linguistic explanations, DRAMA lays a foundation for enhancing perception and communication systems in autonomous vehicles. DRAMA excels in localizing and explaining risks, associating important objects like vehicles or pedestrians with potential hazards and pairing these insights with natural language captions. It benchmarks multitask models that simultaneously identify risks and generate explanations, fostering advancements in vision-language integration. However, DRAMA's natural language annotations focus on assessing and understanding risks, whereas doScenes is built around driving instructions.

C. Bridging Human-Robot Instruction and Autonomous Driving

Unlike previously-mentioned datasets which emphasize scene understanding and description, our research is the first public real-world dataset to provide driving instructions and referentiality information as the natural language annotation, creating a link between imperative language and motion for autonomous vehicles. This task is relevant for autonomous vehicles due to sudden changes to the environment which can create novel or anomalous scenarios, even in spatial locations which may have been typical in moments prior. Vision-language models have been successful in detecting such scene changes [15]. Such anomalies may require a manual takeover response [16], [17], [18], but in cases where a driver is unable to operate the vehicle, the ability of the vehicle to autonomously respond to commands can be especially valuable; this task of navigation of a novel environment without a map but with natural language instruction is often referred to as Vision-and-Language Navigation (VLN) [19], [20]. The task of translating language to actuated action is complex, requiring reasoning, closed-loop planning, and control. NaVILA

[19] addresses this task by adding an intermediate mid-level, language-based representation of actions above low-level control signals, providing a reasonable intermediate in the process of translating language to action. This decoupling of language and control allows one VLA to modularly fit a new robot with appropriate adjustment of low-level control policy, particularly useful towards an autonomous driving setting where different vehicles may have different control configurations. NaviLLM [21] frames the task as a VQA, with an answer set comprised of scene views from multiple directions; the predictive task is to identify which direction the robot should move next toward the completing of a text-prompted goal, evaluating based on whether (and how soon) the robot reaches the desired goal. Essentially, this evaluation is a step-through of panoramic options, which is infeasible for open-world driving where additional agents affect safe reachability of spatial states. With similar one-to-one framing of waypoints as images observed by the robot, LM-Nav [22] uses a goal-conditioned model to infer a sequence of graph vertices to traverse for the purpose of visiting landmarks identified by an LLM and grounded to the scene [23] by a VLM. This approach requires that the robot first collects image and GPS observations over the intended search area. This is, however, less practical for a dynamic environment where landmarks may be other agents themselves, and have no grounding to a map, something reflected in the dynamic-referential instruction subset of doScenes. The algorithm of Hu et al. [24] may better interact with such dynamism by grounding constraint-based instructions through detection-informed adjustments to a costmap which is then used for collision-avoidant navigation. As in additional works, goals are grounded to locations in a predefined semantic map for global motion plans.

III. DOSCENES DATASET

The process of collecting and annotating driving-instruction data is a complex task; a test vehicle must be properly equipped with appropriate sensors for perception, and interior microphones must capture and synchronize verbal commands to driving events. Following collection, expensive annotation of objects and visual scene features must occur to enable supervised learning. Fortunately, massive datasets such as nuScenes provide completion of large portion of this task (that is, large-scale collection and vision-based annotation). We apply retroactive annotation of driving instructions by playing back each of the 1,000 12-second nuScenes clips, and transcribing an instruction (or lack of instruction) that would be given to a driver from the vantage of the passenger to initiate the motion plan observed in the clip.

This natural language instruction can be generated by a heuristic we name the *taxi test*: if you were being driven through this scene by a taxi driver, what instruction, if any, would you need to give an instruction to trigger the behavior observed in the video?

For each of the 1,000 scenes in nuScenes, we provide a set of instruction annotations. These annotations are generated by five independent annotators, and each annotator may include multiple annotations for a scene if they imagine multiple instructions which may generate a similar series of events.



Fig. 2. Histogram of number of instruction annotations per scene; most of the scenes of doScenes have only one or two annotations. Having a greater number of instruction annotations reflects an annotator's generation of multiple possible instructions that could cause the same scene playout.

 TABLE II

 Statistics on Referentiality of doScenes Instructions

Non-Referential535Static Referential214Dynamic Referential159Both93

An instruction field may be blank if no instruction interaction is needed to 'cause' the action (e.g. waiting at a red light, continuing in your lane with the flow of traffic, etc.). Referring to the taxi test, instructions should instigate a change from default vehicle motion.

In addition to the annotated instruction, we provide an additional column for instruction referentiality. When an instruction refers to dynamic objects, e.g. "follow the white van", the *dynamic reference* tag is given. When an instruction refers to static objects, e.g. "stop at the blue sign", the *static reference* tag is given. It is possible for an instruction to be annotated with zero, one, or both of these tags.

We note that even though an instruction may not be referential, it is still expected that the autonomous vehicle (or driver) is fully aware of the major static and dynamic objects in the scene at all times. This is a prerequisite for safe autonomous driving, independent of instruction interactions. Rather, the dynamic referentiality tag is intended to indicate which instructions may require further observation of an object than is available at the moment of instruction. Such instructions cannot be evaluated in an "open loop" manner, since the playout of the scene's dynamic objects will influence the ego motion plan.

IV. CONSIDERATIONS FOR APPLICATION AND EVALUATION

In this section, we provide some observations on the nature of the annotated instructions and their relationship to nuScenes data, with possible implications in how this data may be useful for learning, and where some limitations may exist. While the instructions in doScenes provide information about where a vehicle should move, it does not necessarily provide information about how the vehicle should move, e.g., driving speed or style. This is primarily due to the retroactive annotation approach applied; in future datasets, instructions on driving style or speed may be provided to instigate particular driving responses. Accordingly, motion plans generated by natural language learned from doScenes may not be responsive to prompts related to speed or style.

Further, the 12-second duration of the scenes in the nuScenes dataset, especially when taken at free-flowing urban traffic speeds, may present motion plans longer than a single instruction can cover. This should be accounted for when using doScenes as a basis for evaluation; accurate response to a prompt may be reflected in only the first t seconds of a nuScenes path before the instruction becomes irrelevant after a significant change of scenery or transition to later stages of a multi-step motion plan. This also opens for future research the consideration of frameworks for multi-stage motion planning using natural language.

We chose to create the tags for static and dynamic referential instructions so that motion plans and associated models can be trained or evaluated over particular sets (e.g. those without reference to certain types of objects). For example, models which use only the rasterized map as input, which have found decent success in prior trajectory prediction tasks [25], [26], [27], may be able to learn appropriate motion plans for non-referential instructions, but would be limited without the LiDAR or front-view image as input for making sense of object references.

doScenes was designed to provide data for systems to learn a relationship between instructions and vehicle motion. If such a model can be learned, future research may include techniques to use this model to generate a trajectory based on natural language, or assign a natural language descriptor to a vehicle trajectory, contributing to the task of interpretable, interactive autonomous vehicle motion planning.

As an example of models which may be extended from vision-language to vision-language-action based on prompted instruction, SpatialRGPT [28] learns representations at the instance level (rather than global level) from 3D scene graphs and integrates depth information to enhance VLMs' spatial perception and reasoning capabilities. Importantly, nuScenes provides 3D input in the form of LiDAR point clouds, making it an appropriate dataset for this VLM, and the annotations of doScenes create possibilities for learning information about the corresponding actions to instructional inputs, which can be explored as future research enabled by this dataset, crossing from the generalized robotics domain to the specific challenges of autonomous driving.

V. CONCLUDING REMARKS AND FUTURE RESEARCH

In addition to the application areas for future research identified throughout the paper, in this section, we would like to highlight future research potential for data collection beyond doScenes. doScenes is a novel form of autonomous driving data where natural language instructions are paired with driving scenes and respective sensor time series. However, the instructions in this case are annotated retroactively; while the annotators give their best estimate of an instruction that would have caused such a scene to unfold, this is only a proxy for a true signal. Future data collection should pair true human instructions with action responses. There are a variety of settings (both naturalistic and experimental) which can allow for such collection, and it is reasonable to expect that a higher volume of such high-quality data will enable better learning of corresponding VLA models.

ACKNOWLEDGMENT

The authors thank Mi³ lab members for their contribution to the annotations of the dataset.

REFERENCES

- M. M. Trivedi, T. Gandhi, and J. McCall, "Looking-in and looking-out of a vehicle: Computer-vision-based enhanced vehicle safety," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 1, pp. 108–120, 2007.
- [2] J. Pearl, "Causal diagrams for empirical research," *Biometrika*, vol. 82, no. 4, pp. 669–688, 1995.
- [3] S. Shrestha, Y. Zha, S. Banagiri, G. Gao, Y. Aloimonos, and C. Fermuller, "Natsgd: A dataset with speech, gestures, and demonstrations for robot learning in natural human-robot interaction," 2024.
- [4] H. Walke, K. Black, A. Lee, M. J. Kim, M. Du, C. Zheng, T. Zhao, P. Hansen-Estruch, Q. Vuong, A. He, V. Myers, K. Fang, C. Finn, and S. Levine, "Bridgedata v2: A dataset for robot learning at scale," 2024.
- [5] Y. Wan, J. Mao, and J. B. Tenenbaum, "Handmethat: Human-robot communication in physical and social environments," 2023.
- [6] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," 2020.
- [7] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 4542–4550, 2024.
- [8] Y. Inoue, Y. Yada, K. Tanahashi, and Y. Yamaguchi, "Nuscenes-mqa: Integrated evaluation of captions and qa for autonomous driving datasets using markup annotations," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 930–938, 2024.
- [9] E. Sachdeva, N. Agarwal, S. Chundi, S. Roelofs, J. Li, M. Kochenderfer, C. Choi, and B. Dariush, "Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7513–7522, 2024.
- [10] J. Mao, Y. Qian, J. Ye, H. Zhao, and Y. Wang, "Gpt-driver: Learning to drive with gpt," arXiv preprint arXiv:2310.01415, 2023.
- [11] W. Wang, J. Xie, C. Hu, H. Zou, J. Fan, W. Tong, Y. Wen, S. Wu, H. Deng, Z. Li, *et al.*, "Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving," *arXiv* preprint arXiv:2312.09245, 2023.
- [12] H. Shao, Y. Hu, L. Wang, G. Song, S. L. Waslander, Y. Liu, and H. Li, "Lmdrive: Closed-loop end-to-end driving with large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15120–15130, 2024.
- [13] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, and H. Zhao, "Drivegpt4: Interpretable end-to-end autonomous driving via large language model," *IEEE Robotics and Automation Letters*, 2024.
- [14] S. Malla, C. Choi, I. Dwivedi, J. H. Cho, and J. Li, "Drama: Joint risk localization and captioning in driving," 2023.
- [15] R. Greer and M. Trivedi, "Towards explainable, safe autonomous driving with language embeddings for novelty identification and active learning: Framework and experimental analysis with real-world data sets," *arXiv* preprint arXiv:2402.07320, 2024.
- [16] A. Rangesh, N. Deo, R. Greer, P. Gunaratne, and M. M. Trivedi, "Autonomous vehicles that alert humans to take-over controls: Modeling with real-world data," in 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), pp. 231–236, IEEE, 2021.

- [17] R. Greer, N. Deo, A. Rangesh, M. Trivedi, and P. Gunaratne, "Safe control transitions: Machine vision based observable readiness index and data-driven takeover time prediction," in 27th International Technical Conference on the Enhanced Safety of Vehicles (ESV) National Highway Traffic Safety Administration, no. 23-0331, 2023.
- [18] A. Rangesh, N. Deo, R. Greer, P. Gunaratne, and M. M. Trivedi, "Predicting take-over time for autonomous driving with real-world data: Robust data augmentation, models, and evaluation," *arXiv preprint* arXiv:2107.12932, 2021.
- [19] A.-C. Cheng, Y. Ji, Z. Yang, X. Zou, J. Kautz, E. Bıyık, H. Yin, S. Liu, and X. Wang, "Navila: Legged robot vision-language-action model for navigation," arXiv preprint arXiv:2412.04453, 2024.
- [20] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proceedings of the IEEE conference on computer* vision and pattern recognition, pp. 3674–3683, 2018.
- [21] D. Zheng, S. Huang, L. Zhao, Y. Zhong, and L. Wang, "Towards learning a generalist model for embodied navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13624–13634, 2024.
- [22] 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, pp. 492–504, PMLR, 2023.
- [23] Q. Xu, Y. Hong, Y. Zhang, W. Chi, and L. Sun, "Grounding language to natural human-robot interaction in robot navigation tasks," in 2021 IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 352–357, IEEE, 2021.
- [24] Z. Hu, J. Pan, T. Fan, R. Yang, and D. Manocha, "Safe navigation with human instructions in complex scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 753–760, 2019.
- [25] N. Deo, E. Wolff, and O. Beijborn, "Multimodal trajectory prediction conditioned on lane-graph traversals," in *Conference on Robot Learning*, pp. 203–212, PMLR, 2022.
- [26] T. Gilles, S. Sabatini, D. Tsishkou, B. Stanciulescu, and F. Moutarde, "Thomas: Trajectory heatmap output with learned multi-agent sampling," in *International Conference on Learning Representations*, 2022.
- [27] R. Greer, N. Deo, and M. Trivedi, "Trajectory prediction in autonomous driving with a lane heading auxiliary loss," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4907–4914, 2021.
- [28] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, and S. Liu, "Spatialrgpt: Grounded spatial reasoning in vision-language models," in *NeurIPS*, 2024.