IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Probing Multimodal LLMs as World Models for Driving

Shiva Sreeram[1], Tsun-Hsuan Wang[1], Alaa Maalouf[2], Guy Rosman[3], Sertac Karaman[4], and Daniela Rus[1]

***Abstract*—We provide a sober look at the application of Multimodal Large Language Models (MLLMs) in autonomous driving, challenging common assumptions about their ability to interpret dynamic driving scenarios. Despite advances in models like GPT-4o, their performance in complex driving environments remains largely unexplored. Our experimental study assesses various MLLMs as world models using in-car camera perspectives and reveals that while these models excel at interpreting individual images, they struggle to synthesize coherent narratives across frames, leading to considerable inaccuracies in understanding (i) ego vehicle dynamics, (ii) interactions with other road actors, (iii) trajectory planning, and (iv) open-set scene reasoning. We introduce the EVAL-LLM-DRIVE dataset and DRIVESIM simulator to enhance our evaluation, highlighting gaps in current MLLM capabilities and the need for improved models in dynamic real-world environments.**

***Index Terms*—Performance evaluation and benchmarking, data sets for robotic vision, autonomous vehicle navigation**

I. INTRODUCTION

IN the rapidly evolving field of artificial intelligence, Multimodal Large Language Models (MLLMs), such as GPT-4o [1], have demonstrated unprecedented capabilities in understanding and generating image/text-based content [2]. Recently, MLLMs have been introduced to the realms of driving to improve context understanding [3], extract spatial features from frames to teach a driving policy based on said features, improve the generalization ability of autonomous driving [4], infer system requirements from in-cabin users' commands to meet their intent [5], understand the driving environment [6], and more [7]. However, the performance of these powerful models has not been tested for **scene** (sequence of images) reasoning in a dynamic driving context, let alone one in a closed control loop, and thus, remains an intriguing area of exploration. We ask the question:

- *"Can MLLMs operate as driving world models"?*

Our contribution. To this end, in this work, we study the reasoning capabilities of MLLMs within driving scenarios,

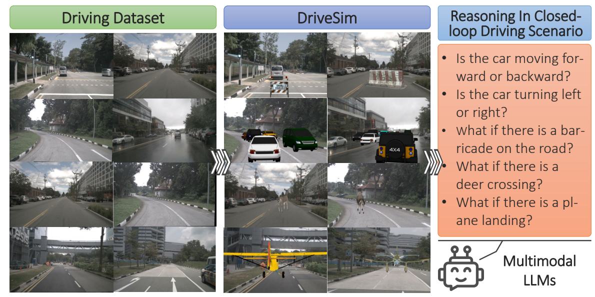

Fig. 1. **Are MLLMs world models for driving?** We investigate their effectiveness in understanding and reasoning about dynamic driving scenarios from sequential images with an introduced real-world driving and re-simulated dataset. Our experiments show that MLLMs struggle to form coherent narratives, failing to reason about car motion, traffic, etc.

aiming to measure their applicability in understanding complex, *dynamic* environments in a variety of scenarios, and their ability to take appropriate actions in decision-making through the integration of *a sequence* of visual data captured from a fixed camera mounted on a driving car *as if the MLLM was the driver*. Specifically, we offer:

- A comprehensive experimental study to evaluate leading MLLMs in their ability to understand and make decisions in dynamic driving scenarios, involving both real driving footage and closed-loop controlled driving. Our tests cover multiple facets of environmental interactions: ego-car dynamics, interactions with other road actors, trajectory planning, and open-set driving scene reasoning. Surprisingly, our findings reveal that MLLMs struggle with interpreting/reasoning and taking correct actions in dynamic driving scenes with significant inaccuracies and biases.
- "EVAL-LLM-DRIVE"[1]: A new dataset designed to provide an array of driving scenarios for evaluating the capabilities of MLLMs in understanding and reasoning about real-world driving scenes from a fixed in-car camera perspective, the same as the driver viewpoint. Real footage captured on the road is the basis of this data, alongside a closed-loop controlled driving simulator DRIVESIM to generate additional diversity and open-set scenarios to the dataset.

A glimpse into our findings. Our experimental results reveal a paradox in the performance of MLLMs. While these models excel at understanding individual images, they struggle to synthesize a coherent narrative across sequences depicting dynamic behavior. This is especially evident in their difficulty in reasoning about vehicular motions, such as identifying whether the ego-car is moving forward or backward. This is headlined by the fact that GPT-4V predicted all scenes as forward-moving, a trend seen in 75.8% of cases with GPT-4o as well! This may stem from a bias in training data, where vehicles predominantly move forward. In planning, these models and others (including Claude3, LLaVA-1.6,

Manuscript received: March, 2, 2025; Revised June, 20, 2025; Accepted August, 22, 2025.

This paper was recommended for publication by Editor Ashis Banerjee upon evaluation of the Associate Editor and Reviewers' comments. This work is supported by Toyota Research Institute (TRI). It, however, reflects solely the opinions and conclusions of its authors and not TRI or any other Toyota entity.

[1]Shiva Sreeram, Tsun-Hsuan Wang, and Daniela Rus are with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139 USA sasreera@mit.edu; tsunw@mit.edu; rus@csail.mit.edu

[2]Alaa Maalouf is with the Department of Computer Science, University of Haifa, Haifa, 3498838, Israel alaamalouf12@gmail.com

[3]Guy Rosman is with the Toyota Research Institute, Cambridge, MA 02139 USA guy.rosman@tri.global

[4]Sertac Karaman is with the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA 02139 USA sertac@mit.edu

Digital Object Identifier (DOI): see top of this page.

[1]The dataset is available at https://github.com/sreeramsa/DriveSim.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

InstructBLIP, and more) consistently failed. When focusing on ChatGPT, it is clear that improvements have been made since the legacy of the GPT-4V model to the latest GPT-4o model, particularly in identifying the dynamic interplay of other road actors. However, there are still failures in other aspects, such as in ego-car dynamics, that prevent it from holding the status as a driving world model. While this shows a positive trajectory in the development of these models, the results highlight additional areas of improvement for the top models. In summary, the experiments highlight a critical gap in the models' ability to connect discrete visual information over time to infer motion, suggesting a limitation in their current state when it comes to understanding the fluidity and continuity inherent in real-world dynamics.

II. RELATED WORK

Lately, the move toward combining different modalities into single large-scale models has gained momentum, such as CLIP [8], BLIP [9], GPT-4V [10] and others [11].

MLLMs in robotics. Recent advancements in robotics have integrated MLLMs, demonstrating their proficiency in engaging effectively within dynamic open-set environments, such as for constructing 3D maps [12], [13], in control and planning [14]–[17], in understanding 3D scenes [18], [19], and in systems for detection and tracking [20]–[22]. Additionally, these models have shown broad adaptability over multi-modal data [19], [20], [23], [24], leading to a new phase where robots can make wise decisions and interact with their surroundings.

MLLMs for driving. In driving, explainable and language-driven representations have gained attention for introspection and event analysis [25]–[29]. Integrating MLLMs into autonomous vehicles enhances vehicle intelligence and user interaction [30] by leveraging real-time data (e.g., traffic, weather) to improve awareness [31] and navigation [32]. LLMs facilitate user-friendly communication for planning [33], [34] and personalize driving settings [35]. They also improve generalization and explainability [4], enhance context awareness [3], interpret user commands [5], and better understand driving environments [6].

Simulation in driving. Training and evaluation of robotic controllers via simulation have become a dominant approach, as evidenced by [36], [37], specifically, in driving [38]–[41]. However, even these simulated environments do not have all of the following features: they (i) can not fully encapsulate the range of vehicle dynamics (e.g., moving forward/backward, turning left/right, etc.), (ii) lack support for adding dynamic characters to scenes for generating interesting driving behaviors (e.g., speeding cars and traffic), and (iii) most importantly they are unlabeled, making it difficult for evaluating MLLMs as world models for driving.

III. PROBING FROM A DATA PERSPECTIVE

Ultimately, a driving world model should encompass multiple facets of environmental interactions and scene reasoning, which we define and test as follows: (i) **Ego-car dynamics:** We check the models' ability to grasp fundamental driving dynamics, such as directionality (forward/backward), velocity changes (acceleration/deceleration), and road adjustments (turning right/left), requiring an understanding in geometric and temporal aspects. (ii) **Dynamic interplay of other road actors:** Progressing beyond the basics, we then challenge the models to reason about the dynamic interplay of other road actors: detecting fast-moving vehicles and discerning traffic jams. (iii) **Planning ability:** Then, we examined the ability of the models to plan accurate driving trajectories, checking whether they can effectively reason the means to avoid obstacles. (iv) **Open-set scene reasoning:** The true test of adaptability lies in open-set reasoning, where our testing challenges conventional driving expectations by creating unforeseen scenarios as unpredictable as airplanes landing on roads or sudden animal appearances, pushing the boundaries of what MLLMs can anticipate and interact correctly with in the meticulously crafted world model. These multi-layer testing scenarios challenge the models' interpretability and decision-making, offering insights into MLLMs' contributions to real-world applications, from alerting wrong-ego car behavior to enhancing navigation systems with real-time alerts on open-set scenes, to traffic updates, and driving validation and planning.

A. Providing the Means to Evaluate a Driving World Model

Surprisingly, the evaluation of MLLMs' scene reasoning in the context of dynamic driving scenarios in the closed-loop control setting remained largely unexplored, potentially due to the lack of a suitable dataset or simulator. To address this gap, we introduce EVAL-LLM-DRIVE; a dataset designed to test MLLMs as driving world models, focusing on the components of a world model and evaluations (i)–(iv). The dataset is divided into two parts: (i) real road footage capturing vehicle dynamics and interactions with other road users, valuable for testing MLLMs reasoning in real scenes, and (ii) scenes generated by DRIVESIM using closed-loop sensor synthesis for simulation[2] (Sec. III-C). We expand upon the real footage with additional diversity through direct manipulation of ego vehicle dynamics and other actors. Furthermore, the simulator enables generation of various open-set scenes involving characters like animals, barriers, and vehicles that are infeasible to obtain in everyday conditions, yet enhance the platform's utility for probing.

B. Scenarios directly from the real world

We drove a 2019 Lexus RX 450H with a 30Hz BFS-PGE-23S3C-CS RGB camera to collect data in Cambridge, MA, and the surrounding area. For most cases, the footage consists of 50% daytime footage, 30% evening/dusk, and 20% night in standard weather conditions. This was all collected in a manner that ensured following rules-of-the-road and safe driving practices, described as follows. **Acceleration/deceleration** captured scenarios involve the ego car legally speeding up (e.g., entering a highway) or slowing down (e.g., approaching a stop sign or turn). **Turns** include both standard intersection turns and subtle curves angling left or right. **Forward** videos show standard driving ahead. Due to traffic rules, **backward** videos are created by reversing forward footage that contains no other moving actors like pedestrians or vehicles. Note that these mimic scenarios such as reversing after overextending at a traffic light intersection without any risk to ourselves or other

[2]Note sensor data was used for generating simulated data and not probing.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I
OVERALL ACCURACY. EVALUATING MLLM PERFORMANCE IN DRIVING TASKS REQUIRING REASONING.

Model	Ego Motion									Other Vehicles						Open-set Reasoning		Logic Problems		
	Forward/Backward			Accelerate/Decelerate			Left/Right			Speeding			Traffic			Object/Animal	Plane Landing	Planning	Behavior	Multi. Causal Inf.
	Real	Sim	Both	Real	Sim	Both	Real	Sim	Both	Real	Sim	Both	Real	Sim	Both	Sim	Sim	Sim	Real + Sim	Real
MiniGPT4-v2	0.50	0.50	0.50	0.50	0.50	0.50	0.55	0.42	0.48	0.50	0.50	0.50	0.50	0.50	0.50	0.20	0.67	0.30	0.67	0.07
InstructBLIP	0.50	0.50	0.50	0.43	0.47	0.45	0.50	0.50	0.50	0.50	0.50	0.50	0.57	**0.73**	0.65	0.25	0.50	0.30	0.73	0.03
LLaVA-1.6	0.48	0.47	0.48	0.50	0.47	0.48	0.47	0.53	0.50	0.50	0.50	0.50	0.68	0.53	0.61	0.27	0.50	0.25	0.76	0.43
GPT-4V	0.50	0.50	0.50	**0.57**	**0.55**	**0.56**	0.60	0.48	0.54	0.52	0.55	0.53	0.63	0.62	0.63	**0.80**	0.63	0.40	-	-
Claude3	0.55	0.48	0.52	0.48	0.47	0.48	0.52	0.50	0.51	0.50	0.55	0.53	0.43	0.62	0.53	0.70	0.53	**0.45**	0.73	0.67
GPT-4o	**0.60**	**0.52**	**0.56**	0.55	0.40	0.48	**0.62**	**0.65**	**0.63**	**0.72**	**0.77**	**0.74**	**0.80**	**0.73**	**0.77**	0.73	**0.70**	**0.45**	**0.94**	**0.97**

drivers. **Traffic** is denoted as situations where traffic causes the ego vehicle to slow down, notably not due to a traffic light but from natural road congestion. Since nighttime traffic is less common, this data has a higher proportion of daytime footage (70% day, 30% evening) compared to other splits. The real footage does not involve as much bumper-to-bumper traffic as we later show can be generated by the simulator due to appropriately safe stopping distances, but still requires the ego vehicle to slow down. Finally, a **"speeding" vehicle** is present when another vehicle overtakes the ego vehicle, and through prompting, the ego is clarified to be driving at the speed limit, implying the overtaking vehicle is speeding.

C. Scenarios by re-simulation of real-world data

Closed-loop sensor synthesis and control. To meet the requirements of our experimental setup, which necessitates a controlled environment and counterfactual testing (as in generating counterfactual data different from the original dataset as opposed to counterfactual reasoning of MLLMs), we develop a data-driven simulator on top of the nuScenes dataset [39]. This approach effectively balances sensor realism [42], [43], closed-loop simulation [38], [44], and scenario setup controllability [45], [46], making it an ideal match for our use case. In the subsequent sections, we outline the key features of the simulator and elucidate their significance to our empirical study for comprehending the reasoning processes of MLLMs within driving scenarios.

Object and actor synthesis in the scene. Building on the described 3D reconstruction pipeline, we seamlessly integrate 3D meshes of desired objects and actors into the scene. These meshes can be efficiently sourced from Objaverse dataset [47] by leveraging the textual comprehension abilities of LLMs on their annotations. For instance, we can identify annotations suggesting that the corresponding meshes represent animals. Utilizing the map's geometric and semantic information, we strategically position the meshes in plausible locations and orientations. Examples include beside the same lane as the ego car, beneath the traffic light, etc.

Behavior modeling of actors. For the behavior of ground vehicles, we employ a Proportional-Integral-Derivative (PID) controller [48] [49] for steering control to track a reference path derived from either the map or a motion plan; for acceleration control, we use an Intelligent Driver Model (IDM) [50] focused on the nearest actor ahead of the ego car moving in a direction potentially in collision with the ego car. For motion planning, we deploy a state lattice planner with quintic polynomial trajectory generation [51], in which the target state lattice is determined to be a specific distance ahead of the ego car in the local frame of its current lane or adjacent lanes. For behavior modeling of other actors, we create trajectories through spline interpolation from predefined start to end poses. Our focus is on modeling the behavior of synthetic actors in reaction to the ego car, to themselves, and to other pre-existing actors/objects in the scene, rather than behaviors of those already existing entities.

IV. EXPERIMENTAL STUDY

Methodology. We used the paradigm explained in Sec. III to test state of the art (SOTA) MLLMs' abilities to determine ego-car motion: (1) is the car proceeding forward or backward? (2) is it accelerating or decelerating? (3) is it turning left or right? All in a categorical manner. Then, we evaluate their reasoning capabilities on other factors in the street to determine whether it detects a speeding car; (4) is there a speeding car?, or heavy traffic; (5) is there heavy traffic? Additionally, we test the decision-making of MLLMs based on an open-set environment by generating open-set scenes by DRIVESIM, such as providing images with the sudden appearance of an animal or static object and even a plane landing: (6) can the ego-car keep moving in the same lane?. We finally can test the capabilities of MLLMs to pick the best trajectory in navigating around obstacles while trying to remain in the lane: (7) which trajectory is the best to follow?

Representing a video scene. We aimed to provide video input to the models to replicate the camera view in real-world driving scenarios. From video input, we create a grid [52], [53] of video frames where each frame is half a second apart, and of resolution 399×224. We test a varied number of frames: three, six, and nine. The total grid resolution is $1259 \times 244/599/712$ for the respective grids, including all frames and white spaces to split them. This format was utilized to avoid concerns with models parsing images in multi-query approaches and avoiding context length limits while **providing a high number of frames**. Another approach involves utilizing a list of base64-encoded frames; however, this approach is **not universal** and in our testing with GPT-4V, the model was **not consistent** in identifying the state of the environment for one frame versus another. Note that while additional sensor data, such as LiDAR or radar, could lead to improvements to accuracy through depth map data, most models don't have this modality and scenes are represented as the aforementioned grids. This setup is showcased in Fig. 4.

Dataset. For each question from (1)–(5), we experiment with EVAL-LLM-DRIVE comprised of two sets, the first set is of real, on-the-road footage (with details including time of day in Sec. III) collected by the authors, and the second set is generated by DRIVESIM. For question (6) we rely on DRIVESIM to generate open set scenes that are hard to capture in real streets. For each of these questions ((1)–(6)), we collect 20 videos in simulation (so for these binary cases there are 10

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

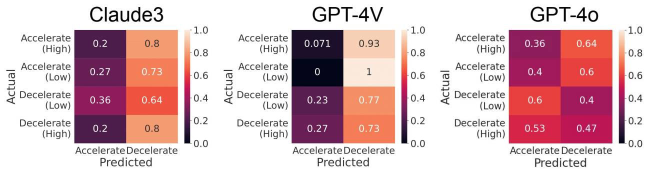

Fig. 2. Accelerate vs decelerate: Confusion matrices.

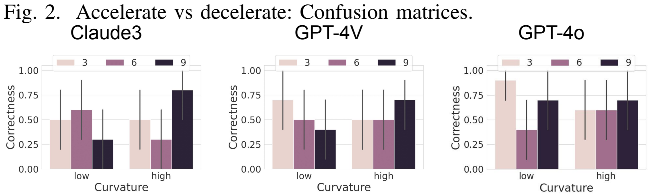

Fig. 3. Left vs right: Performance bar plots.

of each type); since each video is parsed in a grid of 3, 6, and 9, the total number of grids of a question is 60 in simulation and an additional 60 in real footage (similarly from 20 videos) for questions (1)–(5). For question (7), we use DRIVESIM to generate trajectory choices over the road surface and static objects, in five scenes and four scenarios per scene, leading to 20 datapoints.

Prompting. Alongside frames, we must provide an appropriate prompt which informs the model of the format of the image, that the frames come from a camera fixed on a moving car, and obtain a relevant response. This is shown in an example prompt in Fig. 4 which queries whether the ego vehicle is experiencing traffic or no traffic. We follow a similar format when prompting for ego actions and the other actors behavior scenario. We ask to describe what is likely going on in each frame to guarantee the model understands it is parsing a video and in the correct frame order, we can then manually validate its explanation. Additional components asked the model to "guess" and answer "in the following format" to prevent scenarios where the model reports that it is not confident enough to answer. The prompts' phrasing was carefully formulated after experimentation with the models' responses, to ensure that the models accurately processed the frames in the correct sequence order, recognized that the footage was from a camera fixed on a moving car, and provided clear, relevant responses to the questions.

Evaluation and reported results. We compare the obtained results from the MLLM to the ground truth created by the authors. Table I reports how GPT-4o, Claude3, GPT-4V (being GPT-4 Legacy model as of Sep. 2024), LLaVA-1.6, InstructBLIP, and MiniGPT4-v2 perform in these cases. To further expand on the results of this evaluation process given in Table I, we delve into specifics for ego-motion, other actor behavior, open-set, and planning reasoning. We provide a more sophisticated analysis focusing on GPT-4o, GPT-4V (a snapshot of which from March 2024), and Claude3 due to their nature as some of the largest models available and the higher levels of reasoning observed in our evaluation, in addition to tracking the improvements of the top of the line GPT-4 model from March compared to September 2024.

A. Ego Motion Reasoning

Acceleration vs deceleration. In simulator datasets, DRIVESIM generates scenarios where we traverse through the same scene at different rates of acceleration or deceleration, both at high and low rates. The human eye can determine *acceleration/deceleration* by using a reference point in view and observe the change in distance over time, so let us explore the models' capability to do so. As shown in Table I, for most models the performance was roughly 50% across all data, if not worse. To further understand the behavior, we explore the performance of Claude3, a snapshot of GPT-4V from March 2024, and GPT-4o in Fig. 2 on the synthesized data as it provides us the flexibility to generate and verify several acceleration/deceleration rates. We present the confusion matrices for scenarios of high and low acceleration and deceleration rates to the models. Note that in these tests, GPT-4V (March) was biased towards a response of *decelerate*. Interestingly, there were more cases where the model predicted *accelerate* when the ground truth was *decelerate* compared to when the ground truth was *accelerate*. For Claude3, while the bias towards *decelerate* remained, it was less extreme than for GPT-4V. In tracking the improvements from GPT-4V to -4o, we see that this bias is largely removed despite the 48% accuracy overall.

Overall, these results demonstrate a past bias in MLLMs where they frequently responded with decelerating. Although GPT-4o has largely eliminated this, its overall accuracy decreased, so the results still reveal limitations in reasoning about acceleration, a crucial aspect of driving dynamics.

Left vs right. For the simulated set, we stress the models by evaluating the inputs while the ego vehicle curves towards the left or right while progressing at a constant speed, as opposed to turning into a perpendicular lane covered in real-driving footage (which also includes following the road curves). We test with high and low curvature levels. The human eye can determine turning direction by the perspective shift of static elements of the scene, moving from the center to the edge of view over time. Let us explore the models' capabilities.

Table I again shows models near the 50% mark. Real-world clips fare slightly better than sim—expected, since many include a perpendicular turn rather than a gentle bend. By varying curvature in DRIVESIM (Fig. 3), we see accuracy depends on both curvature and clip length. GPT-4V (March) prefers short clips for low curvature but more frames for high curvature. Claude 3 peaks with nine-frame, high-curvature clips and bottoms out with nine-frame, low-curvature ones. The pattern implies that extra temporal context helps when the path bends sharply yet muddies subtler shifts. As for why this would occur, it is likely because with fewer frames, the model can focus more on the curvature of the road to see the gradual turn but with more frames, force it to track inter-frame relations—an MLLM weakness as we've seen. GPT-4o exhibits the trend less strongly—the six-frame, low-curvature case its main outlier.

Forward vs. backward motion comparison is intuitive for humans, relying on basic geometric and dynamic understanding

TABLE II
PERCENTAGE OF *forward / backward* RESPONSES.

	MiniGPT4	InstructBLIP	LLaVA	GPT-4V	Claude 3	GPT-4o
Forward	100%	100%	96.7%	100%	86.7%	75.8%
Backward	0%	0%	2.5%	0%	13.3%	24.2%

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

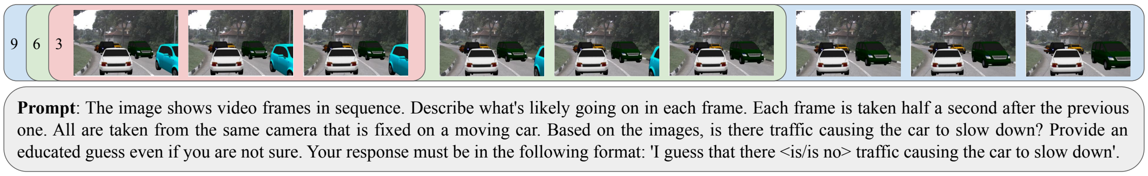

Fig. 4. Heavy traffic scene provided by DRIVESIM converted into a grid (either 1×3, 2×3, or 3×3 respectively) alongside the text prompt.

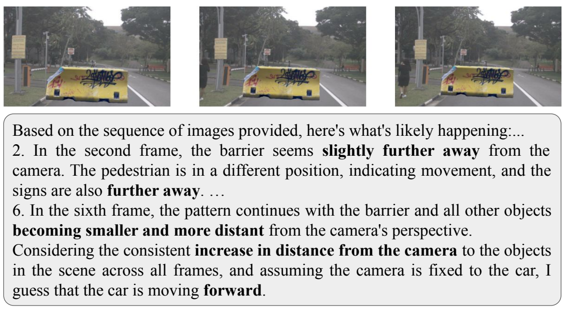

Fig. 5. Frames of backward motion, with an added object, provided by DRIVESIM, alongside truncated GPT-4V response in analyzing the video.

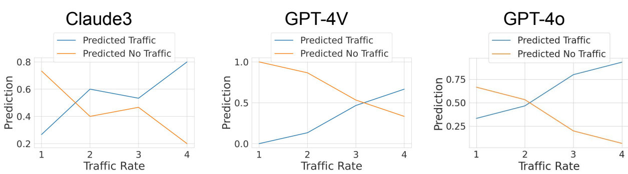

Fig. 6. Line plots depicting a confusion matrix for model performance in simulated traffic vs no traffic scenes.

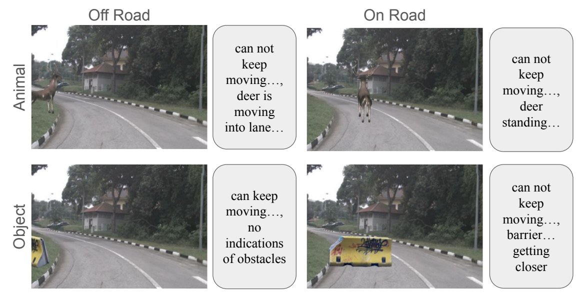

Fig. 7. Object/animal scenario frames with truncated GPT-4o responses.

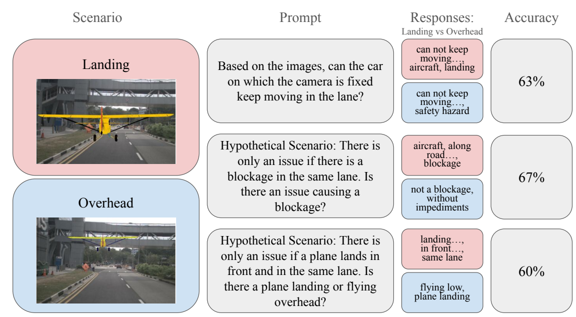

Fig. 8. Plane scenarios with truncated prompts and GPT-4V Legacy responses with accuracy.

(e.g., are objects moving closer or farther?). This test examines models' geometric reasoning, using simulated data at varying constant speeds, similar to other ego-motion experiments. We evaluate the models' understanding in this case. Table I shows the results here were once again roughly 50% for all models. However, there is an extreme underlying bias in the model responses. We further analyze the performance in Table II. This table showcases the percentage of responses from the models that were either *forward* or *backward* regardless of the ground truth. Despite the differences in the scenes, GPT-4V (both the March 2024 snapshot and the Legacy model as of Sep. 2024), alongside most other models, **always reported that the ego vehicle is progressing forward**, with both Claude3 and GPT-4o having a strong forward bias. The bias here is at a level where it overwhelms the evidence and the model responds that if a car is on the road, it must be driving forward as technically, that is what the car should be doing. As such, using DRIVESIM, we probed GPT-4V's failure mode by adding a notable barrier while the ego car reversed away from it. In the three frames of Fig. 5 the barrier steadily recedes, a cue the model correctly notices. Yet, despite recognizing this motion, it still classifies the ego vehicle as moving *forward*. The model has the capability of reasoning required to understand the geometry of the world, yet its responses are biased to such a level that it fails to make accurate predictions.

B. Other Actor Behavior Reasoning

Traffic vs no traffic. There are two main traffic sources: the amount of other vehicles on the road and the speed at which the ego vehicle can move given the other vehicles. As such, geometrical understanding is necessary to witness the number of other vehicles in the scene and the combination of geometric and temporal reasoning for the speed of traffic flow. In real-footage, we have no control over the other vehicles, so traffic is fully binary, but with our simulator, we provide four traffic levels (evenly distributed). (i) The lowest, with label of *no traffic*, is where there is no other vehicle in the same lane. (ii) The second lowest, also labeled *no traffic*, is defined by another vehicle being in the same lane but moving at a high enough speed such that the ego's speed is not hindered. (iii) The next, labeled as *traffic*, is defined by a large number of other vehicles, with slow yet steady traffic flow. (iv) The highest, labeled as *traffic*, involves a large number of actors all moving at a very slow speed, shown in Fig. 4 where to the human eye, the traffic level is clear.

To clarify the concept of "traffic" and remove any subjectivity in interpretation, we must clearly define it within our prompts. Specifically, we refer to traffic as any situation where external factors are causing the vehicle to reduce speed. We operationalize this in the model prompts with the following explicit query: "Is there traffic causing the car to slow down?". This precise definition ensures consistency across evaluations. This case also presents a clear example of how GPT-4o is an improvement, achieving 77% accuracy and thus, a higher level

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

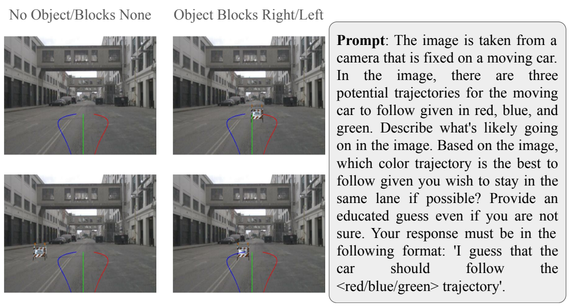

Fig. 9. Images depicting plans with different static object placements and the text prompt.

of reasoning. We explore further results, notably showcasing the changes from GPT-4V (March) to the latest GPT-4o, presented in Fig. 6, illustrating the models' performance using a confusion matrix that captures predictions across various traffic levels. Overall, the prediction positively correlates to traffic level where Claude 3 excels at heavy-traffic cases, while GPT-4V (March) remains biased toward no traffic. We see the marked improvement in identifying traffic in GPT-4o, able to shake what may have been a *no traffic* bias. Despite remaining imperfections, traffic identification is the most reliable capability across ego-motion and other actor scenarios, as reflected in Table I.

Speeding vehicle vs no speeding vehicle identification is critical for road safety. This requires an understanding of geometry and time for perceiving the motion of another actor. Here, we provide scenarios from real-footage where the ego vehicle is overtaken by another vehicle or not and from DRIVESIM, involving another actor that is speeding or not speeding at two-speed levels each. To emphasize this, we must clarify that the ego vehicle is driving at the speed limit such that the relative speed of the other agent can be observed accurately. Given the formulation, a large change in distance between the ego car and the other vehicle should indicate speeding and be sufficient for a human to understand.

As previously, we must remove ambiguity in the query, adjusting the prompt shown in Fig. 4 to ensure the model is aware that the ego vehicle is driving at the speed limit. Therefore, we make the following key change: "All are taken from the same camera that is fixed on a moving car going at the speed limit". When observing the results in Table I, we see a roughly 50% accuracy in the models. We further analyze these results, looking at the percentage of responses from the models that were either *speeding* or *no speeding* regardless of the ground truth. Despite the contrast in the motions, GPT-4V barely detected a speeding vehicle, reporting there was no speeding vehicle 94% of the time across all scenes, a trend shared by the smaller models. Claude3 has a strong bias towards reporting speeding at 87.5%. This level of extreme bias was not present in GPT-4o, where instead the model is the clear front runner. DRIVESIM lets us stress-test these biases: inserting a reference vehicle traveling alongside the ego car—while keeping the prompt unchanged—enables GPT-4V to identify the speeding car, implying its miss stems from weak ego-centric spatial reasoning rather than motion priors. GPT-4o, by contrast, shows no such limitation.

C. Open-Set Reasoning

The seemingly random placement of **animals and static objects** in a scene is one of the open-set scenarios DRIVESIM enables to evaluate MLLMs. As human drivers, we handle these situations instinctively: slow down or avoid static objects on the road, while ignoring those off-road. For animals, uncertainty leads to slowing down or avoiding them regardless. The actions are clear in the scenarios shown in Fig 7. We can see that the large models, GPT-4o, GPT-4V, and Claude3, were quite successful in their reasoning for these cases as shown in Table I and as seen by GPT-4o responses in Fig. 7; however, note that the accuracy of GPT-4o is actually lower than GPT-4V Legacy model so in this case, the newer model has not achieved an improvement.

Plane landings vs flying overhead is a fascinating open-set scenario we explore with DRIVESIM. A human driver may not know how to react to such an extreme case but we can observe the MLLMs behavior. The frames in Fig. 8 showcase a scene with the plane landing or flying overhead. The primary prompt, used for Table I, showed that regardless of the plane landing, the model suggested that you can not keep moving due to the risk: a fair response. As such, we explored a few hypothetical scenarios which really test the geometric and temporal understanding of the plane's motion, which is in a completely different axis than other scenarios.

D. Logic Problems

Planning, or event prediction, experiments are presented using DRIVESIM, generating plans and visualizing them in the camera view. We also introduce static objects to assess whether MLLMs can choose paths that navigate around obstacles. In Fig. 9 we show the four ways we ran planning evaluations for a given scene: (1) no object, (2) the object not blocking anything, (3) the object blocking the middle and right trajectories, and (4) the object blocking the middle and left trajectories. Given the aim of staying in the same lane and presented with the three trajectory choices, a human driver would have a clear pick for each example: (1) green, (2) green, (3) blue, and (4) red. To run evaluation, we utilize a different prompt to pick a trajectory in a single image. As such, we used the prompt shown in Fig. 9, where we also specify the objective of staying in the same lane so there is always one correct choice.

We see in Table I that the larger models achieve significantly better accuracy. However, **their success rate is still surprisingly, under** 50%**,** which is not ideal for closed-loop planning. As such, further probing is required to see the source of the limitations. Claude3's performance improved from 45% to 55% by adding "while avoiding obstacles" to the prompt. This addition showcases the failure of the MLLM as a world model to boost accuracy.

Behavior reasoning tests why the ego acts as it does by turning each clip into a four-option multiple-choice query. For deceleration scenes we offer *animal on road*, *animal about to cross*, *roadblock*, or *malfunction*; when the object is off-road we instead ask why travel is safe, adding distractors such as *traffic police*. Traffic clips—real and simulated—replace the animal-in-lane option with *traffic*. Performance is high across models, bar Claude3, which often hallucinates "animal about to cross," a trend continued in the next section. This mirrors

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

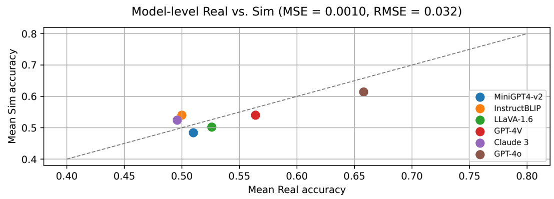

Fig. 10. Visualizing the real vs sim model accuracies with reference line.

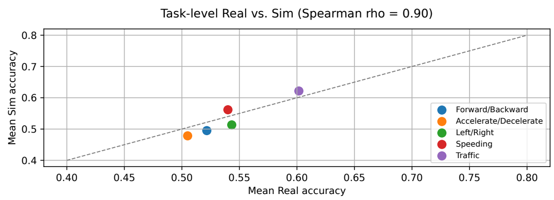

Fig. 11. Visualizing the real vs sim task accuracies with reference line.

their success on other vehicle and open-set tasks where a single frame usually suffices to detect an obstacle or traffic. GPT-4V is omitted because the model has been discontinued.

Multimodal causal inference will have us reason about the reasoning of the ego vehicle. Leveraging our real *speeding* videos, we preface each clip by stating that the ego car has just changed lanes, then pose a multiple-choice query: was the maneuver caused by a speeding car, a roadblock, an animal, or a malfunction? MiniGPT4-v2 and InstructBLIP misfire, habitually selecting "malfunction" and "roadblock," respectively. GPT-4o, however, surpasses its own *speeding* baseline, evidencing strong causal reasoning with this prompt. As before, only a few frames suffice—enough to reveal a fast-moving vehicle and to rule out obstacles/animals.

E. Real versus Simulated Data

Table I shows similar results between real and simulated data, with a few exceptions, like Claude3 in *traffic* tests. However, the model's response patterns remained consistent, as seen in in Table II, and those we discussed when exploring the *speeding* results. Looking at the underlying response strategy for Claude3 in a similar manner to those shown in Table II, we see on real data, the responses for *traffic* and *no traffic* are both 50% where for simulated, the responses were 55% and 45% respectively, so the model was effectively at a random guess level (this can also be seen with the lack of monotonic behavior for Claude3 in Fig. 6). Overall, the root mean squared error between real versus sim across all models is ≈ 0.03 (Fig. 10), and the typical behavior observed in their response strategy is consistent.

Note that much of the driving trajectories closely follow the points of nuScenes and for those that don't (ego motion) we rely on the strengths of the PID controller and IDM discussed in Sec. III. To further quantitatively emphasize the similarity across the five driving maneuvers, the average model accuracy in the sim split closely tracks the real split (Spearman $\rho = 0.90$, p = 0.04), indicating that the simulator preserves the relative task difficulty as shown in Fig. 11.

V. CONCLUSION

This work demonstrates the current capabilities of SOTA MLLMs, including GPT-4o and Claude3, as driving world models. Despite the improvements in GPT-4o over the prior best GPT model alongside strengths in logic problems that require minor spatial-temporal understanding, their limitations in reasoning across multiple frames of driving scenarios have become evident through our extensive experimental results. While many accuracy levels seem random, DRIVESIM allows for probing the reasoning capabilities behind the prediction, exposing details on biases. We observed that failures in handling scenarios stemmed from biases in expected vehicle movements, like assuming forward motion on a road. These results promote future work in delving into the internals of open-source models to understand more about these failures, where more failures could be identified through longer conversations following the responses discovered in this work. In particular, there is potential in the use of spatial-temporal module techniques being researched for video models to use in foundation models, as well as generation and use of richer temporal data that can be used to either train or finetune MLLMs, including with our dataset or an expanded version that includes additional context through peripheral vision, to increase their capabilities.

REFERENCES

[1] OpenAI, "Hello GPT-4o," 2024. [Online]. Available: https://openai.com/index/hello-gpt-4o/

[2] J. Yang, H. Jin, R. Tang, X. Han, Q. Feng, H. Jiang, S. Zhong, B. Yin, and X. Hu, "Harnessing the power of LLMs in practice: A survey on ChatGPT and beyond," *ACM Trans. Knowl. Discov. Data*, feb 2024, just Accepted. [Online]. Available: https://doi.org/10.1145/3649506

[3] L. Chen, O. Sinavski, J. Hünermann, A. Karnsund, A. J. Willmott, D. Birch, D. Maund, and J. Shotton, "Driving with LLMs: Fusing object-level vector modality for explainable autonomous driving," *arXiv preprint arXiv:2310.01957*, 2023.

[4] T.-H. Wang, A. Maalouf, W. Xiao, Y. Ban, A. Amini, G. Rosman, S. Karaman, and D. Rus, "Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

[5] Y. Yang, Q. Zhang, C. Li, D. S. o. Marta, N. Batool, and J. Folkesson, "Human-centric autonomous systems with LLMs for user command reasoning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, January 2024, pp. 988–994.

[6] D. Fu, X. Li, L. Wen, M. Dou, P. Cai, B. Shi, and Y. Qiao, "Drive like a human: Rethinking autonomous driving with large language models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 910–919.

[7] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. Liao *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.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[9] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[10] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[11] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

[12] R. Ding, J. Yang, C. Xue, W. Zhang, S. Bai, and X. Qi, "PLA: Language-driven open-vocabulary 3D scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7010–7019.

[13] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Audio visual language maps for robot navigation," *arXiv preprint arXiv:2303.07522*, 2023.

[14] S. Li, X. Puig, C. Paxton, Y. Du, C. Wang, L. Fan, T. Chen, D.-A. Huang, E. Akyürek, A. Anandkumar *et al.*, "Pre-trained language models for interactive decision-making," *Advances in Neural Information Processing Systems*, vol. 35, pp. 31 199–31 212, 2022.

[15] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "RT-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[16] Y. Bisk, A. Holtzman, J. Thomason, J. Andreas, Y. Bengio, J. Chai, M. Lapata, A. Lazaridou, J. May, A. Nisnevich *et al.*, "Experience grounds language," *arXiv preprint arXiv:2004.10151*, 2020.

[17] S. Tellex, N. Gopalan, H. Kress-Gazit, and C. Matuszek, "Robots that use language," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, pp. 25–55, 2020.

[18] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, T. Funkhouser *et al.*, "Openscene: 3D scene understanding with open vocabularies," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 815–824.

[19] K. M. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, T. Chen, S. Li, G. Iyer, S. Saryazdi, N. Keetha, A. Tewari *et al.*, "Conceptfusion: Open-set multimodal 3D mapping," *arXiv preprint arXiv:2302.07241*, 2023.

[20] A. Maalouf, N. Jadhav, K. M. Jatavallabhula, M. Chahine, D. M. Vogt, R. J. Wood, A. Torralba, and D. Rus, "Follow anything: Open-set detection, tracking, and following in real-time," *IEEE Robotics and Automation Letters*, vol. 9, no. 4, pp. 3283–3290, 2024.

[21] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," *arXiv preprint arXiv:2201.03546*, 2022.

[22] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[23] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[24] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8821–8831.

[25] J. Kim, T. Misu, Y.-T. Chen, A. Tawari, and J. Canny, "Grounding human-to-vehicle advice for self-driving vehicles," Nov. 2019.

[26] S. Tan, B. Ivanovic, X. Weng, M. Pavone, and P. Kraehenbuehl, "Language conditioned traffic generation," *arXiv preprint arXiv:2307.07947*, 2023.

[27] Z. Zhong, D. Rempe, Y. Chen, B. Ivanovic, Y. Cao, D. Xu, M. Pavone, and B. Ray, "Language-guided traffic simulation via scene-level diffusion," *arXiv preprint arXiv:2306.06344*, 2023.

[28] D. Omeiza, H. Webb, M. Jirotka, and L. Kunze, "Explanations in autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 10 142–10 162, 2021.

[29] E. Aasi, P. Nguyen, S. Sreeram, G. Rosman, S. Karaman, and D. Rus, "Generating out-of-distribution scenarios using language models," *arXiv preprint arXiv:2411.16554*, 2024.

[30] C. Cui, Y. Ma, X. Cao, W. Ye, and Z. Wang, "Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 902–909.

[31] ——, "Receive, reason, and react: Drive as you say with large language models in autonomous vehicles," *arXiv preprint arXiv:2310.08034*, 2023.

[32] 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.

[33] M. Omama, P. Inani, P. Paul, S. C. Yellapragada, K. M. Jatavallabhula, S. Chinchali, and M. Krishna, "ALT-Pilot: Autonomous navigation with language augmented topometric maps," *arXiv preprint arXiv:2310.02324*, 2023.

[34] Z. Wang, S. Cai, G. Chen, A. Liu, X. S. Ma, and Y. Liang, "Describe, explain, plan and select: Interactive planning with LLMs enables open-world multi-task agents," in *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 34 153–34 189.

[35] H. Sha, Y. Mu, Y. Jiang, L. Chen, C. Xu, P. Luo, S. E. Li, M. Tomizuka, W. Zhan, and M. Ding, "LanguageMPC: Large language models as decision makers for autonomous driving," *arXiv preprint arXiv:2310.03026*, 2023.

[36] R. Tedrake *et al.*, "Drake: Model-based design and verification for robotics," 2019.

[37] 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.

[38] A. Amini, T.-H. Wang, I. Gilitschenski, W. Schwarting, Z. Liu, S. Han, S. Karaman, and D. Rus, "VISTA 2.0: An open, data-driven simulator for multimodal sensing and policy learning for autonomous vehicles," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2419–2426.

[39] 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," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 621–11 631.

[40] Q. Lu, , X. Wang, Y. Jiang, G. Zhao, M. Ma, and S. Feng, "Multimodal large language model driven scenario testing for autonomous vehicles," https://arxiv.org/pdf/2409.06450v1, 2024.

[41] P. T. Nguyen, T.-H. Wang, Z.-W. Hong, E. Aasi, A. Silva, G. Rosman, S. Karaman, and D. Rus, "Regen: Generative robot simulation via inverse design," in *The Thirteenth International Conference on Learning Representations*, 2025.

[42] Y. Chen, F. Rong, S. Duggal, S. Wang, X. Yan, S. Manivasagam, S. Xue, E. Yumer, and R. Urtasun, "GeoSim: Realistic video simulation via geometry-aware composition for self-driving," in *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 2021, pp. 7230–7240.

[43] Z. Yang, Y. Chen, J. Wang, S. Manivasagam, W.-C. Ma, A. J. Yang, and R. Urtasun, "UniSim: A neural closed-loop sensor simulator," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1389–1399.

[44] S. Suo, S. Regalado, S. Casas, and R. Urtasun, "TrafficSim: Learning to simulate realistic multi-agent behaviors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 400–10 409.

[45] Q. Li, Z. M. Peng, L. Feng, Z. Liu, C. Duan, W. Mo, and B. Zhou, "ScenarioNet: Open-source platform for large-scale traffic scenario simulation and modeling," *Advances in neural information processing systems*, vol. 36, 2024.

[46] W. Ding, Y. Cao, D. Zhao, C. Xiao, and M. Pavone, "RealGen: Retrieval augmented generation for controllable traffic scenarios," *arXiv preprint arXiv:2312.13303*, 2023.

[47] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3D objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 142–13 153.

[48] K. H. Ang, G. Chong, and Y. Li, "PID control system analysis, design, and technology," *IEEE transactions on control systems technology*, vol. 13, no. 4, pp. 559–576, 2005.

[49] K. J. Åström and R. M. Murray, *Feedback Systems: An Introduction for Scientists and Engineers*, 2nd ed. Princeton, NJ: Princeton University Press, 2010. [Online]. Available: https://fbswiki.org/wiki/index.php/Feedback_Systems:_An_Introduction_for_Scientists_and_Engineers

[50] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical review E*, vol. 62, no. 2, p. 1805, 2000.

[51] T. M. Howard, C. J. Green, A. Kelly, and D. Ferguson, "State space sampling of feasible motions for high-performance mobile robot navigation in complex environments," *Journal of Field Robotics*, vol. 25, no. 6-7, pp. 325–345, 2008.

[52] Fusseldieb, "Reading videos with GPT4V," 2023. [Online]. Available: https://community.openai.com/t/reading-videos-with-gpt4v/523568

[53] C. Wan, X. Luo, H. Luo, Z. Cai, Y. Song, Y. Zhao, Y. Bai, F. Wang, Y. He, and Y. Gong, "Grid: Omni visual generation," 2025. [Online]. Available: https://arxiv.org/abs/2412.10718