STSBench: A Spatio-temporal Scenario Benchmark for Multi-modal Large Language Models in Autonomous Driving

Christian Fruhwirth-Reisinger[1,2,*] **Dušan Malić**[1,2,*] **Wei Lin**[3]
David Schinagl[1] **Samuel Schulter**[4,†] **Horst Possegger**[1,2]

[1]Institute of Visual Computing, Graz University of Technology
[2]Christian Doppler Laboratory for Embedded Machine Learning
[3]Institute for Machine Learning, Johannes Kepler University Linz
[4]Amazon

Code: `https://github.com/LRP-IVC/STSBench`
Data: `https://huggingface.co/datasets/ivc-lrp/STSBench`

Abstract

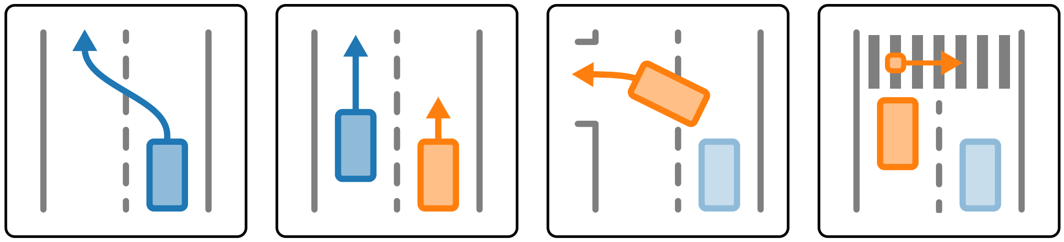
We introduce STSBench, a scenario-based framework to benchmark the holistic understanding of vision-language models (VLMs) for autonomous driving. The framework automatically mines pre-defined traffic scenarios from any dataset using ground-truth annotations, provides an intuitive user interface for efficient human verification, and generates multiple-choice questions for model evaluation. Applied to the NuScenes dataset, we present STSnu, the first benchmark that evaluates the spatio-temporal reasoning capabilities of VLMs based on comprehensive 3D perception. Existing benchmarks typically target off-the-shelf or fine-tuned VLMs for images or videos from a single viewpoint and focus on semantic tasks such as object recognition, dense captioning, risk assessment, or scene understanding. In contrast, STSnu evaluates driving expert VLMs for end-to-end driving, operating on videos from multi-view cameras or LiDAR. It specifically assesses their ability to reason about both ego-vehicle actions and complex interactions among traffic participants, a crucial capability for autonomous vehicles. The benchmark features 43 diverse scenarios spanning multiple views and frames, resulting in 971 human-verified multiple-choice questions. A thorough evaluation uncovers critical shortcomings in existing models' ability to reason about fundamental traffic dynamics in complex environments. These findings highlight the urgent need for architectural advances that explicitly model spatio-temporal reasoning. By addressing a core gap in spatio-temporal evaluation, STSBench enables the development of more robust and explainable VLMs for autonomous driving.

1 Introduction

The rapid development of increasingly powerful vision-language models (VLMs) [4, 8, 9, 11, 32, 33, 38–40] has sparked significant interest in applying them to end-to-end autonomous driving [18, 24, 29, 34, 47, 52, 54, 60, 71, 74, 75]. These models aim to enhance trust in fully autonomous systems by providing human-interpretable decisions in natural language [65]. Unlike pre-trained generalist

[*]Equal contribution. Corresponding authors: `{reisinger, dusan.malic}@tugraz.at`
[†]This work is independent of the author's employment at Amazon.

Preprint.



(a) Ego. (b) Agent-to-Ego. (c) Agent. (d) Agent-to-Agent.

Figure 1: **STSBench scenario categories.** The benchmark covers common ego-vehicle (blue) actions, *e.g.*, *ego lane change* (a) and interactions with agents (orange), *e.g.*, *ego overtaking agent* (b), important for vehicle control. In addition, to test for complex spatio-temporal understanding, we evaluate agent actions, *e.g.*, *agent left turn* (c), and interactions between agents, *e.g.*, *agent waiting for pedestrian to cross* (d).

VLMs, these driving experts operate on consecutive multi-view images or LiDAR scans to understand a scene comprehensively and are fine-tuned for planning and controlling the ego vehicle.

To achieve end-to-end driving, where raw sensor inputs are directly mapped to driving actions, most models [29, 60, 75] predict future trajectories through waypoints or control signals for the ego-vehicle. A common evaluation strategy is to perform open-loop planning on NuScenes [6], as most other real-world planning datasets lack raw visual inputs alongside map data and agent trajectories, or are synthetic benchmarks [27] simulated in CARLA [17]. A drawback of the open-loop evaluation on NuScenes [6], however, is the relatively small and unbalanced validation split [37] in which, for approximately 75% of all cases, the correct action is "continue to drive straight". Moreover, evaluating L2 error between predicted and ground truth waypoints, or measuring the accuracy of ego-action predictions, offers limited insight into the decision-making and reasoning capabilities of language-based end-to-end driving experts. Therefore, further testing of these models is crucial.

A growing number of benchmarks assess the understanding of VLMs in the context of automated driving [20, 35, 42, 44, 51, 53, 56, 65, 69]. They typically focus on specific abilities such as spatial reasoning in camera images [20, 56], recognition and prediction of ego-vehicle actions [44, 53], the handling of visual corruptions [65], or understanding risk [42, 51] and critical driving scenarios [69]. However, most of these benchmarks target general-purpose VLMs that operate on single images or monocular videos and are not designed to evaluate whether models can jointly reason over spatially distributed and temporally extended inputs from multi-view video or LiDAR data, which is an essential capability for coherent understanding in complex, real-world driving scenes. Furthermore, most existing action- or event-based benchmarks evaluate only the behavior of the ego-vehicle. While this focus is reasonable for planning and control tasks, it overlooks the broader situational understanding required for safe driving, such as anticipating interactions between other traffic participants. Even if they are not of immediate importance, driving models should have the ability to understand such actions or the future consequences of events.

Another major challenge lies in developing annotation schemes for benchmarks applicable to existing perception or planning datasets without requiring extensive manual effort. Existing driving-related benchmarks [30, 31, 44, 51, 53, 57, 65] are typically tied to a specific dataset through manual annotations or extensive human verification. However, although many driving datasets [6, 43, 55, 62] are recorded using multi-camera and LiDAR systems, their sensor setups differ significantly regarding camera type, placement, and orientation. As a result, a model with 3D understanding [36, 59] trained and evaluated on one dataset cannot be straightforwardly assessed using a benchmark built for another, since the projection from images to 3D space depends on the dataset-specific calibration. This makes it necessary to create separate benchmark annotations for each dataset, which is both time-consuming and costly when done manually.

To address the identified issues, we introduce STSBench, a generalizable framework for automatically mining spatio-temporal driving scenarios from existing datasets with rich ground-truth annotations. The framework identifies traffic scenarios, following a pre-defined scenario catalog, that reflect real-world interactions by leveraging information such as 3D bounding boxes and tracks, object class labels, ego-motion data, and HD maps. The catalog contains all scenarios and their definitions. It

Table 1: **Task-specific driving benchmarks.** Autonomous driving benchmarks created from the NuScenes [6] dataset focusing on various tasks grouped by dataset source. †: we do not consider simple status classification annotations such as *moving*, *walking*, *etc*. as temporal reasoning. Annotation modalities denote single images (S.I.), multi-view images (M.I.), and multi-view videos (M.V.). Evaluation types are: Visual Question Answering (VQA), Multiple Choice (MC), Numerical (NUM), and Open-loop Planning (OLP).

Benchmark Name	Anno. Modality	Anno. Type	Human Verif.	Spat. Reas.	Temp. Reas.†	Multi-view Events	Third Party Interaction	Eval. Type	Pub. Avail.
DriveMLLM [20]	S.I.	Auto	✓	✓	✗	✗	✗	MC, NUM	✓
NuScenes-MQA [25]	M.I.	Auto	✗	✓	✗	✗	✗	VQA	✓
NuScenes-QA [48]	M.I.	Auto	✗	✓	✗	✗	✗	VQA	✓
DriveLM [53]	M.I.	Manual	✓	✓	✗	✓	✓	VQA	✓
NuScenes-SpatialQA [56]	M.I.	Auto	✗	✓	✗	✗	✗	MC, NUM	✗
DriveBench [65]	M.I.	Manual	✗	✓	✗	✓	✓	VQA	✓
NuInstruct [16]	M.V.	Auto	✗	✓	✓	✓	✗	VQA	✓
TOD3Cap [30]	M.V.	Manual	✓	✓	✗	✗	✗	VQA	✓
DriveLLM-o1 [26]	M.I.	Auto	✓	✓	✓	✗	✗	VQA, MC	✓
OmniDrive [60]	M.V.	Auto	✓	✓	✓	✗	✗	VQA, OLP	✓
STSBench3D (ours)	M.V.	Auto	✓	✓	✓	✓	✓	MC	✓

assigns each scenario a list of negative scenarios that do not occur in the same scene. In addition to this fully automated extraction procedure, we provide a visual inspection tool that enables fast and effortless human verification of the mined scenarios. Inspectors are tasked to check for false positive scenarios and remove negative ones that also apply for the corresponding scenario (*e.g.*, for a vehicle increasing its speed while performing a *right turn*, *accelerate* would be removed). From the verified scenarios, STSBench automatically constructs a multiple-choice benchmark that asks models to identify which interactions occur in a given scene. Questions may concern ego-vehicle behavior, the actions of other agents (vehicles, pedestrians, or cyclists), interactions between agents and the ego-vehicle, or between multiple third-party agents. Examples of the different categories are illustrated in Fig. 1. The workflow is simple: automatically mine scenarios, verify them with minimal effort, and convert them into a structured spatio-temporal reasoning task.

Furthermore, we instantiate STSBench on the validation split of the NuScenes dataset, which remains the most commonly used training and evaluation dataset for vision-language models in autonomous driving. Unlike existing benchmarks that focus narrowly on ego-centric actions in images or monocular videos, our benchmark, STSnu, explicitly targets spatio-temporal reasoning involving both ego and non-ego agents across multiple views and time steps (see Table 1). STSnu comprises 43 scenarios resulting in a total of 971 challenging multiple-choice questions.

We conduct a detailed evaluation of various models that fall into one of three categories: text-only large language models (LLMs), off-the-shelf VLMs, or driving expert VLMs. While LLMs receive ground truth trajectories, off-the-shelf VLMs operate on multiple images. Expert models are designed to deal with consecutive multi-view images. With just trajectory information available, LLMs outperform both VLM counterparts significantly. Our evaluations highlight that state-of-the-art models across all categories provide limited spatio-temporal reasoning capabilities. This is especially notable for more challenging scenarios (involving interactions between other agents), which require a truly holistic understanding of the scene.

2 Related Work

Driving datasets and benchmarks with text annotations. Autonomous driving (AD) is an extensive field of research that has led to the creation of numerous datasets [6, 10, 19, 43, 55, 62, 68] for various perception tasks. These datasets have been enriched with text annotations to facilitate language-based model training for specific tasks in the AD domain. Following common practices in visual instruction tuning [11, 38, 33], annotations have been added mostly for separate multi-view images, focusing on tasks such as grounding [14, 58], ego-action prediction [31, 66], open-loop planning [53, 57], risk assessment [15, 42], spatial reasoning [20, 56], or visual question answering [44, 45, 35]. To address the lack of 3D understanding, EML [76] introduces text annotations that incorporate question-answer pairs about 2D-to-3D spatio-temporal relations. The first attempts to evaluate the spatial capabilities of vision-language models for AD are the benchmarks DriveM-LLM [20], NuScenes-MQA [25], and NuScenes-SpatialQA [56]. They asses VLMs for their ability to measure distances and understand relative positions within camera images. However, since autonomous driving necessitates a holistic understanding of dynamic scenes, various datasets that incorporate multi-view video [16, 60] or 3D [48, 30, 64, 51] annotations emerged. Datasets such as Nuscenes-QA [48], DriveLM [53], OmniDrive [60], and NuInstruct [16] propose visual question-answering frameworks aimed at scene understanding, chain-of-thought reasoning and counterfactual reasoning. Despite their extensive annotations, these datasets rely predominantly on question–answer pairs that emphasize semantics [48, 53] and spatial relations [16], with limited temporal context, particularly beyond ego-vehicle interactions. STSnu specifically targets these gaps and tests the spatio-temporal reasoning capabilities of end-to-end driving models.

Vision-language models for end-to-end driving. Vision-Language Models (VLMs) [4, 9, 13, 2, 33, 39, 11] have attracted a lot of attention due to their exceptional zero-shot capabilities. These capabilities have also raised interest in applying these models to end-to-end AD for open-loop and closed-loop planning. Early methods directly apply VLMs to the front-view camera images of an autonomous vehicle [3, 41, 53, 67, 73, 76] to predict future trajectories or control signals in text form. However, a holistic understanding of the traffic scene is crucial for realistic driving scenes with highly dynamic scenarios. Therefore, another line of work operates on multi-view images and videos [29, 23, 34, 60]. To cope with the increasing number of image tokens caused by additional views and multiple video frames, Senna [29] compresses each view via temporal attention for path planning. Another technique for dealing with multiple views is encoding into Bird's Eye View (BEV) features that are later aligned with the underlying LLM. While GPVL [34] and BEV-InMM [16] utilize a BEVFormer [36] backbone, OmniDrive [60] uses StreamPETR [59]. More recent approaches combine end-to-end driving models [21, 28] with VLMs [18, 47, 54, 75] to increase the contextual understanding and provide reasoning alongside future trajectories. Despite their excellent performance on the planning task, the variety of methods and their handling of available input modalities raise questions about their environmental understanding (*i.e.*, reasoning capabilities). Therefore, we conduct a detailed analysis about the spatio-temporal reasoning capabilities of language-based end-to-end driving models by applying STSBench to instantiate STSnu on NuScenes [6] data.

3 Spatio-temporal Scenario Benchmark

With STSBench, we introduce a benchmark framework designed to evaluate the spatio-temporal reasoning capabilities of vision-language models (VLMs) in autonomous driving. While most existing benchmarks focus on off-the-shelf or fine-tuned VLMs operating on single images or monocular videos, our benchmark targets expert driving models. These experts are expected to have a comprehensive 3D understanding of dynamic scenes and therefore need to process multi-view, LiDAR, or a combined video input signal that enables holistic reasoning. The development of our benchmark is motivated by two observations:

There is a gap in assessing the spatio-temporal understanding of expert driving models. Recent efforts have adapted VLMs for driving [24, 29, 60] or extended existing planning models such as UniAD [21] with LLMs [75] to improve interpretability and trust. However, these expert models are usually evaluated on the NuScenes [6] dataset using predicted waypoints or control signals, with metrics such as L2 error or collision rate. Even if these scores are excellent, they do not guarantee that the model's decisions are grounded in a correct understanding of other traffic participants or scene dynamics. Although several benchmarks have been developed to test off-the-shelf VLMs for spatial [20, 25, 56] or temporal [16, 26] reasoning, most are restricted to single frames or monocular views. NuInstruct [16] remains an exception by providing multi-view video-based questions, but it is automatically generated and lacks human verification, making it better suited for training than evaluation. Additionally, the temporal reasoning aspect of the NuInstruct benchmark is limited to rather simple motion states, such as whether an agent is moving or stopped.

Existing benchmarks are tailored for a specific dataset and can hardly be transferred or extended by additional scenarios. Encoding a 3D scene from multiple 2D images implicitly requires knowledge of the transformations that relate each image to a shared 3D space. Most recent detection encoders [36, 59], commonly used in end-to-end driving models, learn this mapping

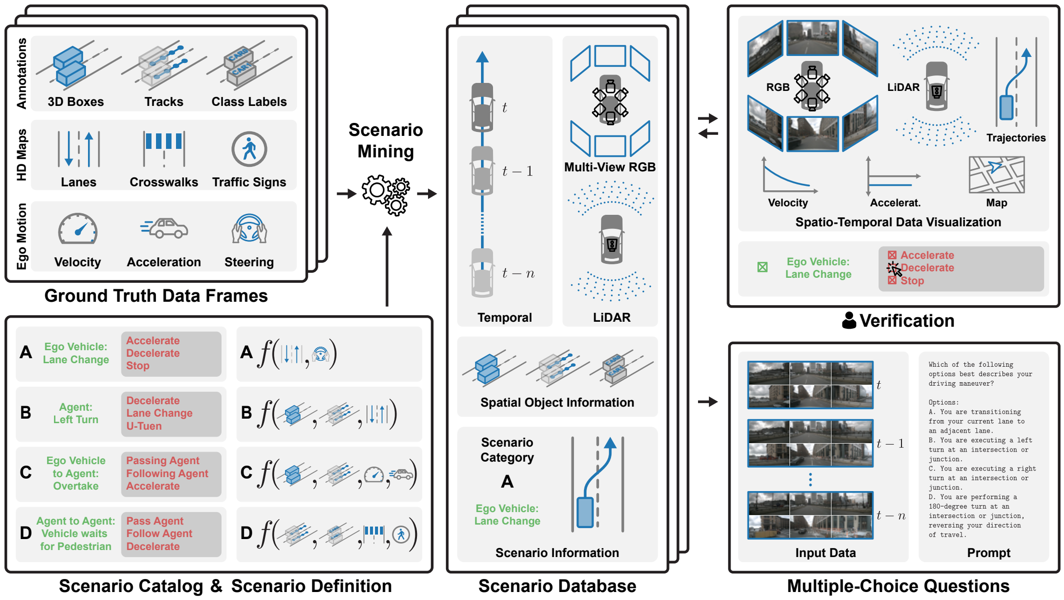


Figure 2: **STSBench workflow.** (Best viewed on screen)

during training. As a result, they become tied to a fixed sensor setup for which they can reason accurately about object sizes and distances. This reliance on specific camera setups makes it difficult to evaluate a model trained on one dataset using the benchmark of another, unless domain adaptation or generalization strategies are employed. Therefore, it is essential to have a fast and generalizable framework applicable to various datasets with different sensor configurations. Moreover, existing benchmarks are often built by manually annotating data [30, 53], making them time-consuming and expensive to replicate. Even when using the same dataset, adding a new benchmark task typically requires inspecting and re-annotating a large number of samples. These challenges underscore the need for a fast and generalizable framework for benchmark creation.

To address these challenges, we introduce STSBench, a framework for fast and scalable traffic scenario mining and verification. In Section 3.1, we describe the annotation pipeline that transforms raw dataset annotations into structured multiple-choice evaluations for spatio-temporal reasoning. Section 3.2 details the instantiation of this framework on the NuScenes [6] dataset, resulting in our STSnu benchmark.

3.1 Annotation Framework

Our annotation framework aims to create an accurate benchmark for any given perception dataset (*e.g.*, [6, 43, 55]), considering all available ground truth annotations for the respective dataset. These annotations include 3D bounding boxes, tracking identifiers, object class labels, ego-motion data (*e.g.*, velocity, acceleration, steering angle), or HD map data (*e.g.*, lanes, lane boundaries, speed limits, crosswalks). In Fig. 2, we illustrate the benchmark creation workflow of STSBench.

1) Scenario catalog. To build a coherent benchmark, we define a scenario catalog containing all relevant scenarios (*e.g.*, lane change, overtake, following, *etc.*). In addition, we assign negative scenarios for each entry of the catalog. Negatives are other scenarios that do not occur in the corresponding scenario. Assume we have an *overtaking* scenario in which a vehicle in motion passes another vehicle, also in motion, in the adjacent lane. Closely related scenarios would be *passing*, where only one of the two vehicles is moving, and *acceleration*, where a vehicle increases its speed. For an actual *overtaking* scenario, both *passing* and *acceleration* are valuable negatives to test the reasoning capabilities: Holistic spatiotemporal understanding implies that also closely related scenarios can be distinguished.

2) Scenario definition. In order to enable automatic mining, we define heuristics for the traffic scenarios we want to test, taking into account available ground truth annotations. For instance, an important scenario crucial for driving experts is recognizing *lane changes* for the ego-vehicle and other vehicles in the scene. The complexity of this action requires knowing the position of the corresponding vehicle throughout multiple frames w.r.t. lane boundaries. The event gets recognized as such if the boundary is crossed and a vehicle transitions from one lane to another.

3) Scenario mining. We automatically mine the pre-defined scenarios and save them in a scenario database. The database contains consecutive images of all available views, LiDAR data, the spatial coordinates of all objects involved in the scenario, and the extracted scenario information, including the found scenario and assigned negatives.

4) Verification. Human verification is used to ensure annotation quality. Rather than reviewing full sequences frame by frame, the annotator only needs to perform two simple checks: confirm or reject the presence of a mined scenario (false positive) and verify that the negative examples are truly invalid (false negatives). For instance, in an *overtaking* scenario, a model might also detect *acceleration*, which could be an invalid negative. These checks are lightweight and fast, allowing for efficient quality control without the burden of traditional manual annotation.

5) Question generation. Finally, STSBench generates multiple-choice questions asking which of the provided scenario examples occur in the scene. For different scenario types, *i.e.*, ego, agent, ego-to-agent, and agent-to-agent, we provide fixed questions containing the required spatial positions of the occurring objects. The questions are designed to have one correct answer and, by default, provide five possible choices. Further details are provided in Appendix A.4.

3.2 STSnu Benchmark Construction

We leverage STSBench to mine scenarios and subsequently derive multiple-choice questions from a real-world dataset for evaluating spatio-temporal reasoning capabilities of end-to-end driving models.

Data Source. Since most expert driving models [22, 30, 57, 60] operate on the multi-view videos or LiDAR scans of NuScenes [6], we construct our benchmark on this large-scale autonomous driving dataset with rich 3D annotations in a multi-sensor setup. In particular, we automatically gather scenarios from all 150 scenes of the validation set, considering only annotated key frames. Therefore, we leverage manually annotated 3D tracks and class labels, ego-motion data (*e.g.*, velocity) from the inertial measurement unit (IMU), and lanes, lane boundaries, and road markings (*e.g.*, crosswalks) from the available HD map data.

In contrast to prior benchmarks, focusing primarily on ego-vehicle actions that mainly occur in the front-view, STSnu evaluates spatio-temporal reasoning across a broader set of interactions and multiple views. This includes reasoning about other agents and their interactions with the ego-vehicle or with one another. To support this, we define four distinct scenario categories:

1) Ego-vehicle scenarios. The first category includes all actions related exclusively to the ego-vehicle, such as *acceleration/deceleration*, *left/right turn*, or *lane change*. Important for control decisions and collision prevention, driving models must be aware of the ego-vehicle status and behavior. Although these scenarios are part of existing benchmarks in different forms and relatively straightforward to detect, they provide valuable negatives for scenarios with ego-agent interactions.

2) Agent scenarios. Similar to ego-vehicle scenarios, agent scenarios involve a single agent in the scene. However, this category additionally contains vulnerable road users such as pedestrians and cyclists. Pedestrians, contrary to vehicles, perform actions such as *walking*, *running*, or *crossing*. Awareness of other traffic participants and their actions is crucial when it comes to risk assessment, planning the next ego action, or analyzing the situation in a dynamic environment. In contrast to ego-vehicle actions, other road users may be occluded or far away and, therefore, pose a particular challenge.

3) Ego-to-agent scenarios. The third category of scenarios describes ego-related agent actions. Directly influencing the driving behavior of each other, this category is similarly important to the ego-vehicle scenarios w.r.t. the immediate control decisions. Ego-agent scenarios contain maneuvers

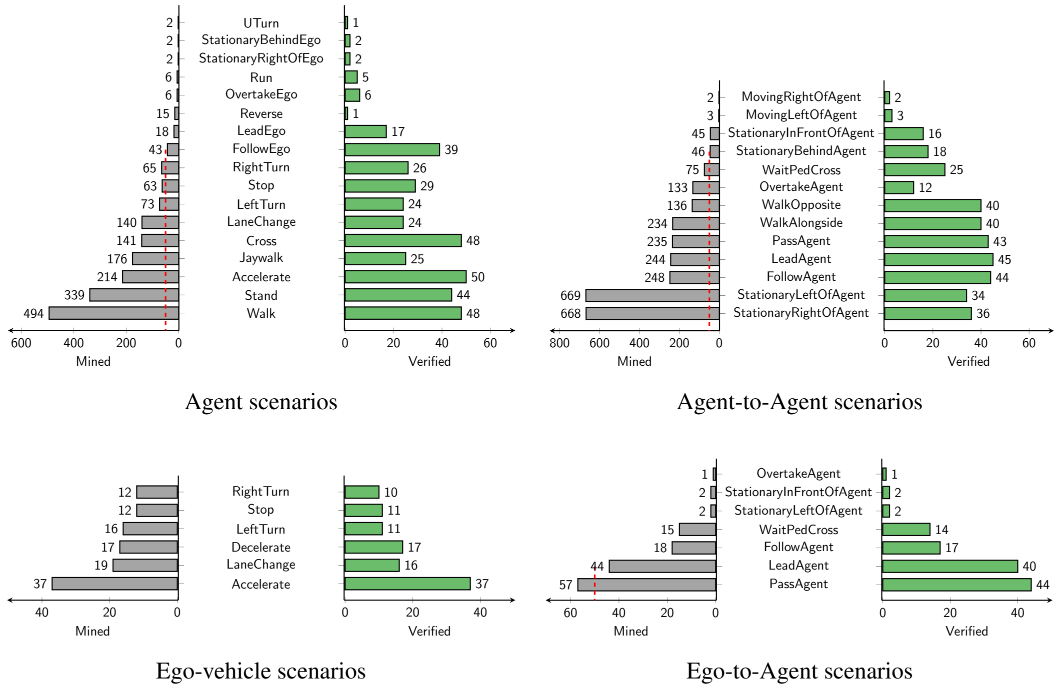


Figure 3: **Scenario statistics.** Number of mined scenarios in total (gray) and the remaining samples (green) after sub-sampling and verification. Scenarios with more than 50 samples (dashed red line) have been sub-sampled considering spatial distribution, occlusion, and distance to the ego-vehicle.

such as *overtaking*, *passing*, *following*, or *leading*. The scenarios focus on agents in the immediate vicinity of the ego-vehicle and direct interactions.

4) Agent-to-agent scenarios. The most challenging group of scenarios concerns interactions between two agents, not considering the ego-vehicle. These scenarios describe the spatio-temporal relationship between objects. For instance, a vehicle that *overtakes* another vehicle in motion or pedestrians *moving alongside* each other. The latter is a perfect example of interactions that do not actively influence the driving behavior of the expert model. However, we argue that a holistic understanding of the scene should not be restricted to the immediate surroundings of the ego-vehicle.

3.3 Benchmark Statistics

The scenario catalog of our STSnu benchmark comprises 43 different scenario descriptions. Using this catalog, STSBench has automatically mined 4790 scenarios from 150 sequences of the NuScenes [6] validation set. To ensure a better balance of the benchmark, we sub-sampled over-represented scenarios based on occlusion rate and spatial distribution of agents. Hence, with this optional step, we removed very difficult examples of highly occluded agents and objects that are far away. The remaining 1188 scenarios have gone through human verification and finally resulted in 971 multiple-choice questions offering at least four possible answers per question, of which exactly one is correct. We provide detailed numbers of mined, sub-sampled, and remaining scenarios after human verification in Fig. 3. The distribution of scenarios assigned to the four proposed categories is illustrated in Fig 4. We can see that a large proportion of the scenarios

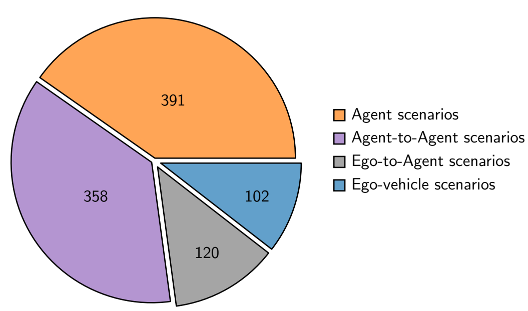


Figure 4: **Scenario distribution.** Number of scenarios per category.

Table 2: **Overall performance.** Performance comparison of LLMs, off-the-shelf VLMs, and driving expert VLMs. Accuracy grouped by scenario categories. The best results are highlighted in bold.

	Ego	Ego-to-Agent	Agent	Agent-to-Agent	Average
Llama 3.2 [1]	16.63	19.73	13.87	21.84	18.01
DeepSeek V3 [12]	31.09	71.49	41.98	44.24	47.20
GPT-4o [46]	**63.63**	**75.75**	45.59	43.38	**57.08**
InternVL 2.5 1B [9]	19.98	38.17	20.47	29.88	27.12
Qwen2.5-VL 7B [5]	35.62	26.64	36.58	37.47	34.07
InternVL 2.5 8B [9]	38.38	40.82	**53.33**	**51.77**	46.07
Senna-VLM [29]	8.81	44.44	26.10	31.45	27.70
OmniDrive [60]	24.40	42.97	23.78	26.15	29.33
DriveMM [22]	43.49	48.31	38.00	28.26	39.51

cover the two more difficult scenario categories, *i.e.*, agent and agent-to-agent. Despite this, the ego and ego-to-agent scenarios, some of which were also examined in other benchmarks, are sufficiently represented. We provide additional statistics and details of STSnu in Appendix A.

Verification and correction. To ensure accuracy, the verification was carried out by three different people with a European class B driver's license. Prior to the validation procedure, the human driving experts were briefed to gain a general understanding of the scenarios and what requirements they need to meet. The agreement on positive samples was 85.6% (1017 agreements out of 1188) while there was a disagreement of 20.8% (247 out of 1188) on negative samples. There were mainly differences in the perception of recognizable or unrecognizable agents and misunderstandings of interactions between two agents. On average, the three inspectors spent 4.78 hours each verifying the data, which is around 14.5 seconds per sample and inspector. A more detailed analysis and verification times are provided in Appendix A.3. For the final STSnu benchmark, we merged positive scenarios applying a majority voting and kept all negatives with full agreement over all reviewers.

4 Experiments

The evaluation on STSnu follows a simple protocol. For each scenario, we measure accuracy as the proportion of correctly answered multiple-choice questions. To account for scenario imbalance, we report the overall accuracy as a weighted mean across all scenarios. Beyond its simplicity, this evaluation method offers the significant advantage of being both interpretable and comparable across models.

Baselines. We evaluate a range of models from three categories: large-language models (LLMs), off-the-shelf vision-language models (VLMs), and driving expert VLMs. First, we task LLMs to infer the correct maneuver given ground-truth perception data. Therefore, we selected two open-source models, DeepSeek V3 [12] and Llama 3.2 [1], and one closed-source model, GPT-4o [46]. Second, for the evaluation of VLMs without fine-tuning, we use Qwen2.5-VL 7B [5], InternVL 2.5 1B [9], and InternVL 2.5 8B [9]. Third, as representative driving expert models, we evaluate Senna-VLM [29], OmniDrive [60], and DriveMM [22].

Evaluation setting. Since only driving expert models are designed to handle multi-view video data, we adapted the input format for LLMs and off-the-shelf VLMs in our evaluation. To simulate a perfect perception system, we provide LLMs with the GPS positions of the ego-vehicle and, when relevant, the trajectories of involved agents, along with the task description, a multiple-choice question, and scenario definitions for the available answers. This setup serves as a simple baseline for comparison. For off-the-shelf VLMs, we supply a series of images and an adapted text prompt. Because scenarios can span multiple viewpoints, we select, at each time step, the image corresponding to the camera view in which the relevant part of the scenario occurs. The associated camera view metadata is also provided. Driving expert models, on the other hand, receive full multi-view image sequences in addition to the text prompt, in which we refer to the involved agents. Further implementation details and input formatting for all models are provided in Appendix E.

Overall performance and analysis. The evaluation results in Table 2 indicate that VLMs, particularly driving expert VLMs, do not have a spatio-temporal understanding of dynamic traffic scenes. Driving expert VLMs are good at basic perception tasks, *i.e.*, observing traffic participants near the vehicle, but struggle with ego-vehicle and agent-to-agent scenarios, thus demonstrating insufficient holistic spatio-temporal understanding. In comparison, without image inputs, the top-performing LLM outperforms its visual counterparts by a significant margin, especially for relatively simple ego-vehicle (63.63% vs. 43.49%) and ego-to-agent (75.75% vs. 48.31%) scenarios. GPT-4o [46], the advanced reasoning model, performs particularly well and reaches an average accuracy of 57.08%. DriveMM [22] is the best performing expert model on average. However, all off-the-shelf VLMs and the top-two LLMs show superior performance in the difficult agent-to-agent scenarios.

An interesting observation is the considerable gap between DriveMM [22] and the other expert models, particularly for the ego scenarios. While OmniDrive projects multi-view image features into BEV, DriveMM directly processes multi-view videos. It is worth noting that the StreamPETR [59] encoder of OmniDrive [60] has initially been designed for perception tasks such as 3D object detection, where it is important to model the surroundings of the ego-vehicle. The results suggest that these representations may hamper the reasoning performance w.r.t. ego actions. We provide a more detailed analysis in Appendix C.

5 Conclusion

In this paper, we introduced STSBench, a framework for automatic scenario mining from large-scale autonomous driving datasets with rich ground truth annotations. The framework also includes a fast and lightweight verification interface, enabling the effortless creation of high-quality benchmarks for spatio-temporal reasoning in multi-view video or LiDAR data. We applied STSBench to the NuScenes dataset resulting in STSnu, which comprises 971 multiple-choice questions derived from 43 diverse traffic scenarios. This benchmark provides a rigorous evaluation of driving expert models in terms of their spatio-temporal understanding from a holistic, scene-level perspective. Our evaluation revealed that, despite recent progress, current driving expert models still show significant limitations in spatio-temporal reasoning, highlighting the need for further research in this area.

Limitations. The major issue with publicly available large-scale driving datasets is that they mostly have clean and homogeneous data collection and filtering processes. For instance, NuScenes has been recorded in Boston and Singapore, and contains hardly unexpected or dangerous driving behavior. This is perfectly fine for perception tasks and also desirable from a human safety perspective. However, this is disadvantageous for the automatic mining of diverse traffic scenarios. Nevertheless, evaluations on available data already pose significant challenges for driving expert models regarding their spatio-temporal reasoning capabilities. Another bottleneck of automatic scenario mining is the careful design of heuristics. For example, spatio-temporal processes have variable length. For instance, the time frame to detect *u-turns* is significantly longer than, for example, *lane changes*, especially when the agent has to interrupt the maneuver because of the current traffic situation. However, considering the rich annotations available, a simple set of heuristics can already provide sufficient pre-selection of traffic scenarios.

Social impact. Our work contributes to safe automated or autonomous driving systems: With STSnu, we highlight the lack of holistic scene understanding of state-of-the-art models. To mitigate the limitations (*e.g.* potential geographic bias, cannot test for safety-critical driving behavior due to lack of such data), we also open-source STSBench, a framework to easily extract and (most importantly) efficiently verify such benchmarks from other datasets. We explicitly rely on heuristics to ensure that the extracted benchmark scenarios are deterministic, easily reproducible and intuitive. We believe that our framework is a valuable and easy-to-use tool to guide future research on driving expert models towards better holistic scene understanding capabilities, in order to achieve safe and trustworthy systems.

References

[1] Meta AI. LLaMA 3.2: Open Foundation and Instruction Models, 2024.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. In *NeurIPS*, 2022.

[3] Hidehisa Arai, Keita Miwa, Kento Sasaki, Yu Yamaguchi, Kohei Watanabe, Shunsuke Aoki, and Issei Yamamoto. CoVLA: Comprehensive Vision-Language-Action Dataset for Autonomous Driving. In *WACV*, 2024.

[4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025.

[6] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[7] Xu Cao, Tong Zhou, Yunsheng Ma, Wenqian Ye, Can Cui, Kun Tang, Zhipeng Cao, Kaizhao Liang, Ziran Wang, James Rehg, and Chao Zheng. MAPLM: A Real-World Large-Scale Vision-Language Dataset for Map and Traffic Scene Understanding. In *CVPR*, 2024.

[8] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024.

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024.

[10] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *CVPR*, 2016.

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *NeurIPS*, 2023.

[12] DeepSeek-AI. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[13] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Vision-Language Models, 2024.

[14] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. Talk2Car: Taking control of your self-driving car. In *EMNLP-IJCNLP*, 2019.

[15] Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. HiLM-D: Towards High-Resolution Understanding in Multimodal Large Language Models for Autonomous Driving. *arXiv preprint arXiv:2308.12966*, 2023.

[16] Xinpeng Ding, Jinahua Han, Hang Xu, Xiaodan Laing, Xu Hang, Wei Zhang, and Xiaomeng Li. Holistic Autonomous Driving Understanding by Bird's-Eye-View Injected Multi-Modal Large Models. In *CVPR*, 2024.

[17] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An Open Urban Driving Simulator. In *CoRL*, 2017.

[18] Haoyu Fu, Diankun Zhang, Zongchuang Zhao, Jianfeng Cui, Dingkang Liang, Chong Zhang, Dingyuan Zhang, Hongwei Xie, Bing Wang, and Xiang Bai. ORION: A Holistic End-to-End Autonomous Driving Framework by Vision-Language Instructed Action Generation. *arXiv preprint arXiv:2503.19755*, 2025.

[19] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *CVPR*, 2012.

[20] Xianda Guo, Zhang Ruijun, Duan Yiqun, He Yuhang, Chenming Zhang, and Long Chen. DriveMLLM: A Benchmark for Spatial Understanding with Multimodal Large Language Models in Autonomous Driving. *arXiv preprint arXiv:2411.13112*, 2024.

[21] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented Autonomous Driving. In *CVPR*, 2023.

[22] Zhijian Huang, Chengjian Fen, Feng Yan, Baihui Xiao, Zequn Jie, Yujie Zhong, Xiaodan Liang, and Lin Ma. Drivemm: All-in-one large multimodal model for autonomous driving. *arXiv preprint arXiv:2412.07689*, 2024.

[23] Zhijian Huang, Tao Tang, Shaoxiang Chen, Sihao Lin, Zequn Jie, Lin Ma, Guangrun Wang, and Xiaodan Liang. Making Large Language Models Better Planners with Reasoning-Decision Alignment. In *ECCV*, 2024.

[24] Jyh-Jing Hwang, Runsheng Xu, Hubert Lin, Wei-Chih Hung, Jingwei Ji, Kristy Choi, Di Huang, Tong He, Paul Covington, Benjamin Sapp, Yin Zhou, James Guo, Dragomir Anguelov, and Mingxing Tan. Emma: End-to-end multimodal model for autonomous driving. *arXiv preprint arXiv:2410.23262*, 2024.

[25] Yuichi Inoue, Yuki Yada, Kotaro Tanahashi, and Yu Yamaguchi. NuScenes-MQA: Integrated Evaluation of Captions and QA for Autonomous Driving Datasets Using Markup Annotations. In *WACVW*, 2024.

[26] Ayesha Ishaq, Jean Lahoud, Ketan More, Omkar Thawakar, Ritesh Thawkar, Dinura Dissanayake, Noor Ahsan, Yuhao Li, Fahad Shahbaz Khan, Hisham Cholakkal, Ivan Laptev, Rao Muhammad Anwer, and Salman Khan. Drivelmm-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding. *arXiv preprint arXiv:2503.10621*, 2025.

[27] Xiaosong Jia, Zhenjie Yang, Qifeng Li, Zhiyuan Zhang, and Junchi Yan. Bench2Drive: Towards Multi-Ability Benchmarking of Closed-Loop End-To-End Autonomous Driving. In *NeurIPS*, 2024.

[28] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. VAD: Vectorized Scene Representation for Efficient Autonomous Driving. In *ICCV*, 2023.

[29] Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging Large Vision-Language Models and End-to-End Autonomous Driving. *arXiv preprint arXiv:2410.22313*, 2024.

[30] Bu Jin, Yupeng Zheng, Pengfei Li, Weize Li, Yuhang Zheng, Sujie Hu, Xinyu Liu, Jinwei Zhu, Zhijie Yan, Haiyang Sun, Kun Zhan, Peng Jia, Xiaoxiao Long, Yilun Chen, and Hao Zhao. TOD3Cap: Towards 3D Dense Captioning in Outdoor Scenes. In *ECCV*, 2024.

[31] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual Explanations for Self-Driving Vehicles. In *ECCV*, 2018.

[32] Junnan Li, Dongxu Li, Caiming Xiong, and S. Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*, 2022.

[33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023.

[34] Tengpeng Li, Hanli Wang, Xianfei Li, Wenlong Liao, Tao He, and Pai Peng. Generative Planning with 3D-vision Language Pre-training for End-to-End Autonomous Driving. *arXiv preprint arXiv:2501.08861*, 2025.

[35] Yanze Li, Wenhua Zhang, Kai Chen, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, et al. Automated Evaluation of Large Vision-Language Models on Self-driving Corner Cases. In *WACV*, 2025.

[36] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. BEVFormer: Learning Bird's-Eye-View Representation from Multi-Camera Images via Spatiotemporal Transformers. In *ECCV*, 2022.

[37] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and Jose M. Alvarez. Is Ego Status All You Need for Open-Loop End-to-End Autonomous Driving? . In *CVPR*, 2024.

[38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *NeurIPS*, 2023.

[39] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *CVPR*, 2024.

[40] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[41] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal Language Model for Driving. In *ECCV*, 2024.

[42] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. DRAMA: Joint Risk Localization and Captioning in Driving. In *WACV*, 2023.

[43] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, and Chunjing Xu. One Million Scenes for Autonomous Driving: ONCE Dataset. In *NeurIPS*, 2021.

[44] Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoit Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, and Oleg Sinavski. LingoQA: Video Question Answering for Autonomous Driving. In *ECCV*, 2024.

[45] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2Drive: Towards Interpretable and Chain-based Reasoning for Autonomous Driving. In *ECCV*, 2024.

[46] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2024.

[47] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. VLP: Vision Language Planning for Autonomous Driving. In *CVPR*, 2024.

[48] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. NuScenes-QA: A Multi-Modal Visual Question Answering Benchmark for Autonomous Driving Scenario. In *AAAI*, 2024.

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021.

[50] Rerun Development Team. Rerun: A Visualization SDK for Multimodal Data, 2024. URL `https://www.rerun.io`. Available from https://www.rerun.io/ and https://github.com/rerun-io/rerun.

[51] Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel Kochenderfer, Chiho Choi, and Behzad Dariush. Rank2Tell: A Multimodal Driving Dataset for Joint Importance Ranking and Reasoning. In *WACV*, 2024.

[52] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L. Waslander, Yu Liu, and Hongsheng Li. LMDrive: Closed-Loop End-to-End Driving with Large Language Models. In *CVPR*, 2024.

[53] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with Graph Visual Question Answering. In *ECCV*, 2024.

[54] Ruiqi Song, Xianda Guo, Hangbin Wu, Qinggong Wei, and Long Chen. InsightDrive: Insight Scene Representation for End-to-End Autonomous Driving. *arXiv preprint arXiv:2503.13047*, 2025.

[55] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *CVPR*, 2020.

[56] Kexin Tian, Jingrui Mao, Yunlong Zhang, Jiwan Jiang, Yang Zhou, and Zhengzhong Tu. NuScenes-SpatialQA: A Spatial Understanding and Reasoning Benchmark for Vision-Language Models in Autonomous Driving. *arXiv preprint arXiv:2504.03164*, 2025.

[57] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Zhiyong Zhao, Yang Wang, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. In *CoRL*, 2024.

[58] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Object Referring in Videos With Language and Human Gaze. In *CVPR*, 2018.

[59] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring Object-Centric Temporal Modeling for Efficient Multi-View 3D Object Detection. In *ICCV*, 2023.

[60] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M. Alvarez. OmniDrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *arXiv preprint arXiv:2405.01533*, 2024.

[61] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *NeurIPS*, 2022.

[62] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next Generation Datasets for Self-driving Perception and Forecasting. In *NeurIPS*, 2021.

[63] Katharina Winter, Mark Azer, and Fabian B. Flohr. BEVDriver: Leveraging BEV Maps in LLMs for Robust Closed-Loop Driving. *arXiv preprint arXiv:2503.03074*, 2025.

[64] Dongming Wu, Wencheng Han, Tiancai Wang, Yingfei Liu, Xiangyu Zhang, and Jianbing Shen. Language Prompt for Autonomous Driving. In *AAAI*, 2025.

[65] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are VLMs Ready for Autonomous Driving? An Empirical Study from the Reliability, Data, and Metric Perspectives. *arXiv preprint arXiv:2501.04003*, 2025.

[66] Yiran Xu, Xiaoyin Yang, Lihang Gong, Hsuan-Chu Lin, Tz-Ying Wu, Yunsheng Li, and Nuno Vasconcelos. Explainable Object-Induced Action Decision for Autonomous Vehicles . In *CVPR*, 2020.

[67] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024.

[68] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *CVPR*, June 2020.

[69] Tong Zeng, Longfeng Wu, Liang Shi, Dawei Zhou, and Feng Guo. Are Vision LLMs Road-Ready? A Comprehensive Benchmark for Safety-Critical Driving Video Understanding. *arXiv preprint arXiv:2504.14526*, 2025.

[70] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *ICCV*, October 2023.

[71] Rui Zhao, Qirui Yuan, Jinyu Li, Haofeng Hu, Yun Li, Chengyuan Zheng, and Fei Gao. Sce2DriveX: A Generalized MLLM Framework for Scene-to-Drive Learning, 2025.

[72] Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large Language Models Are Not Robust Multiple Choice Selectors. In *ICLR*, 2024.

[73] Wenzhao Zheng, Zetian Xia, Yuanhui Huang, Sicheng Zuo, Jie Zhou, and Jiwen Lu. Doe-1: Closed-Loop Autonomous Driving with Large World Model. *arXiv preprint arXiv: 2412.09627*, 2024.

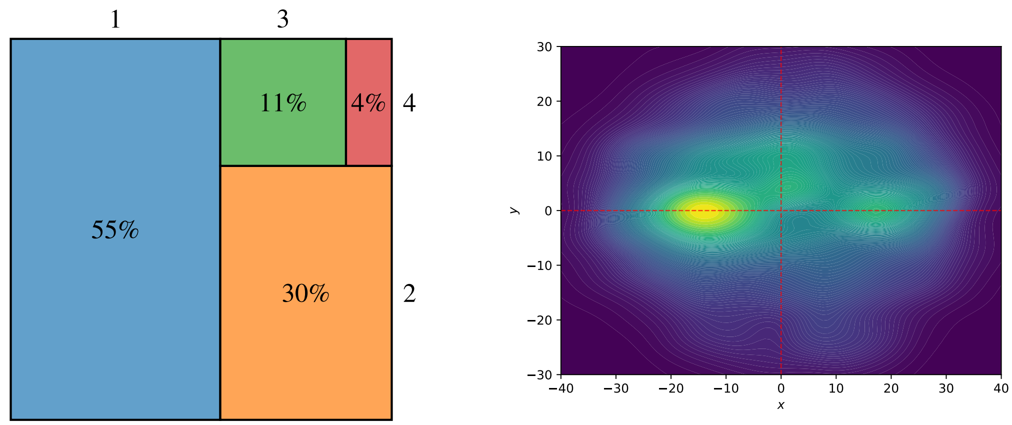
[74] Xin Zhou, Dingkang Liang, Sifan Tu, Xiwu Chen, Yikang Ding, Dingyuan Zhang, Feiyang Tan, Hengshuang Zhao, and Xiang Bai. HERMES: A Unified Self-Driving World Model for Simultaneous 3D Scene Understanding and Generation. *arXiv preprint arXiv:2501.14729*, 2025.

[75] Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, and Alois C. Knoll. OpenDriveVLA: Towards End-to-end Autonomous Driving with Large Vision Language Action Model. *arXiv preprint arXiv:2503.23463*, 2025.

[76] Yunsong Zhou, Linyan Huang, Qingwen Bu, Jia Zeng, Tianyu Li, Hang Qiu, Hongzi Zhu, Minyi Guo, Yu Qiao, and Hongyang Li. Embodied Understanding of Driving Scenarios. In *ECCV*, 2024.

Appendix

Table of Contents



(a) Number of camera views per scenario.

(b) Heatmap of x-y positions of different scenarios.

Figure 5: **Scenario statistics.** Distribution of scenarios with agent involvement across camera views and on the x-y plane.

A Benchmark details

A.1 STSnu statistics

Since driving scenes are dynamic environments in which both the ego-vehicle and other traffic participants are in motion, a single camera view is often insufficient to capture all relevant interactions. As a result, most modern datasets [6, 55, 62] employ multi-camera systems with slightly overlapping fields of view. In our benchmark, we mine scenarios where agents are distributed across varying numbers of camera views, as illustrated in Fig.5a. We observe that approximately half of the scenarios are confined to a single view, while the remaining scenarios span up to four views over the observed time period. The relatively high concentration of actions in a single view can be attributed to two main factors: 1) In the majority of sequences in NuScenes, the ego-vehicle either drives straight or remains stationary. 2) Many scenarios occur in the front or rear view of the ego-vehicle, as these cameras primarily capture the road it travels on. Each of these factors, individually and in combination, contributes to the skew in this distribution. Nevertheless, the remaining half of the scenarios span multiple views, underscoring the importance of evaluating a model's ability to generalize across spatially distributed visual inputs.

To maintain a balanced dataset, we sub-sample the mined scenarios and remove samples from overrepresented categories. We define three criteria for agent-related scenarios to guide this selection: 1) occlusion, 2) distance to the ego-vehicle, and 3) spatial distribution. Based on these criteria, we retain scenarios that are highly visible, occur in the near surrounding of the ego-vehicle, and are spatially well-distributed around it. The first criterion is straightforward since visual systems cannot effectively reason about occluded agents. The second criterion prioritizes agents that are closer to the ego-vehicle, as distant objects are harder to perceive, and the benchmark is not focused on evaluating long-range detection or reasoning about small objects. The final criterion ensures a diverse set of samples with respect to camera viewpoints and spatial coverage, as illustrated in Fig.5b. We observe a distributional imbalance, with a disproportionate number of scenarios occurring in the rear area of the ego-vehicle, reflecting biases in the underlying data. However, the remaining scenarios are relatively well distributed across other spatial regions.

A.2 Scenario catalog

The benchmark generation starts with the definition of a scenario catalog. It includes all scenarios that should be covered in the benchmark and assigns negative scenarios for each entry. Negative scenarios do not occur during the assigned scenario and serve later in the benchmark creation as wrong-choice options for the multiple-choice answer generation. In our scenario catalog, we list scenarios for ego-vehicle (Fig. 6), other agents (Fig. 8), interactions between ego-vehicle and agents (Fig. 7), and interactions among other agents (Fig. 9). For the respective categories, we define the scenarios in text form in Tables 5, 4, 3 and, 6.

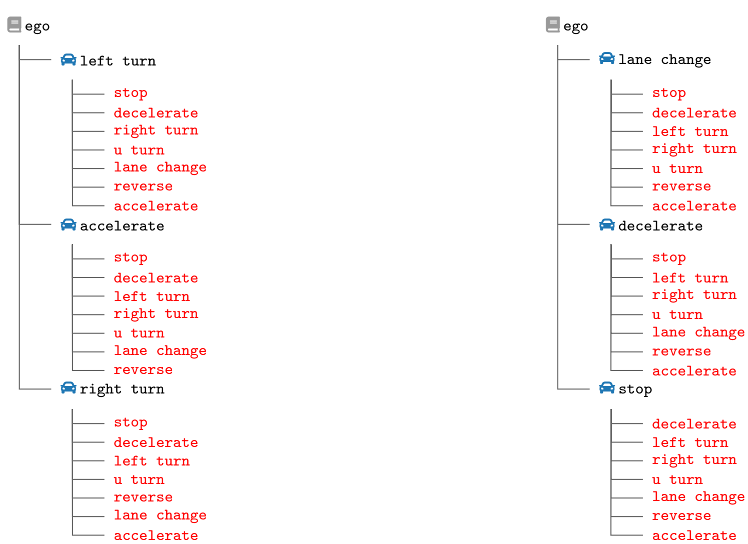


Figure 6: **Scenario catalog.** Ego scenarios with assigned negative scenarios in red.

- ego-agent
  - lead agent
    - stationary behind agent
    - pass agent
    - stationary right of agent
    - stationary in front of agent
    - stop
    - right turn
    - accelerate
    - moving right of agent
    - u turn
    - follow agent
    - stationary left of agent
    - moving left of agent
    - lane change
    - overtake agent
    - reverse
    - wait ped cross
    - decelerate
    - left turn
  - pass agent
    - stationary behind agent
    - stationary right of agent
    - stationary in front of agent
    - stop
    - right turn
    - left turn
    - accelerate
    - moving right of agent
    - u turn
    - follow agent
    - stationary left of agent
    - moving left of agent
    - lane change
    - overtake agent
    - reverse
    - wait ped cross
    - lead agent
    - decelerate
  - wait ped cross
    - pass agent
    - follow ego
    - decelerate
    - reverse
    - u turn
    - overtake ego
    - follow agent
    - lead ego
    - overtake agent
    - right turn
    - lead agent
    - left turn
    - stop
  - follow agent
    - stationary behind agent
    - pass agent
    - stationary right of agent
    - overtake agent
    - stationary in front of agent
    - stop
    - left turn
    - reverse
    - accelerate
    - moving right of agent
    - u turn
    - moving left of agent
    - lane change
    - stationary left of agent
    - right turn
    - wait ped cross
    - lead agent
    - decelerate

- ego-agent
  - overtake agent
    - pass agent
    - moving left of agent
    - moving right of agent
    - stationary left of agent
    - stationary right of agent
    - lead agent
    - follow agent
    - stationary in front of agent
    - stationary behind agent
    - wait ped cross
    - accelerate
    - decelerate
    - stop
    - lane change
    - u turn
    - reverse
    - left turn
    - right turn
  - stationary left of agent
    - pass agent
    - overtake agent
    - moving left of agent
    - moving right of agent
    - stationary right of agent
    - lead agent
    - follow agent
    - stationary in front of agent
    - stationary behind agent
    - wait ped cross
    - accelerate
    - decelerate
    - stop
    - lane change
    - u turn
    - reverse
    - left turn
    - right turn
  - stationary in front of agent
    - lead agent
    - follow agent
    - stationary behind agent
    - pass agent
    - overtake agent
    - moving left of agent
    - moving right of agent
    - stationary left of agent
    - stationary right of agent
    - wait ped cross
    - accelerate
    - decelerate
    - stop
    - lane change
    - u turn
    - reverse
    - left turn
    - right turn

Figure 7: **Scenario catalog.** Ego-Agent scenarios with assigned negative scenarios in red.

- agent
  - right turn
    - stop
    - decelerate
    - left turn
    - u turn
    - lane change
    - reverse
    - accelerate
  - lane change
    - stop
    - decelerate
    - left turn
    - right turn
    - u turn
    - reverse
    - accelerate
  - overtake ego
    - stationary right of ego
    - moving left of ego
    - follow ego
    - stationary in front of ego
    - stop
    - decelerate
    - left turn
    - right turn
    - moving right of ego
    - pass ego
    - u turn
    - stationary left of ego
    - lead ego
    - lane change
    - stationary behind ego
    - reverse
    - accelerate
  - follow ego
    - stationary behind ego
    - lead ego
    - stationary in front of ego
    - pass ego
    - overtake ego
    - stationary left of ego
    - stationary right of ego
    - moving left of ego
    - moving right of ego
    - accelerate
    - lane change
    - left turn
    - right turn
    - u turn
    - reverse
    - stop
    - decelerate
  - stop
    - decelerate
    - left turn
    - right turn
    - u turn
    - lane change
    - reverse
    - accelerate
  - run
    - stop
    - decelerate
    - cross
    - accelerate
    - jaywalk
    - stand
    - walk

- agent
  - lead ego
    - stationary right of ego
    - moving left of ego
    - follow ego
    - stationary in front of ego
    - stop
    - left turn
    - reverse
    - moving right of ego
    - pass ego
    - accelerate
    - u turn
    - overtake ego
    - stationary left of ego
    - lane change
    - stationary behind ego
    - right turn
    - decelerate
  - reverse
    - stop
    - decelerate
    - left turn
    - accelerate
    - u turn
    - right turn
  - u turn
    - decelerate
    - lane change
    - left turn
    - right turn
    - stop
    - reverse
  - stationary right of ego
    - overtake ego
    - pass ego
    - moving right of ego
    - moving left of ego
    - stationary left of ego
    - follow ego
    - lead ego
    - stationary in front of ego
    - stationary behind ego
    - accelerate
    - decelerate
    - lane change
    - left turn
    - right turn
    - u turn
    - reverse
    - stop
  - stationary behind ego
    - follow ego
    - lead ego
    - stationary in front of ego
    - pass ego
    - overtake ego
    - stationary left of ego
    - stationary right of ego
    - moving left of ego
    - moving right of ego
    - accelerate
    - decelerate
    - lane change
    - left turn
    - right turn
    - u turn
    - reverse
    - stop

- agent
  - jaywalk
    - cross
    - run
    - stand
    - accelerate
    - decelerate
    - stop
  - left turn
    - decelerate
    - lane change
    - right turn
    - u turn
    - stop
    - reverse
    - accelerate
  - walk
    - cross
    - jaywalk
    - stand
    - run
    - accelerate
    - decelerate
    - stop
  - stand
    - cross
    - jaywalk
    - walk
    - run
    - accelerate
    - decelerate
    - stop
  - cross
    - jaywalk
    - run
    - stand
    - accelerate
    - decelerate
    - stop
  - accelerate
    - lane change
    - left turn
    - right turn
    - u turn
    - decelerate
    - reverse
    - stop

Figure 8: **Scenario catalog.** Agent scenarios with assigned negative scenarios in red.

- agent-agent
  - follow agent
    - stationary behind agent
    - lead agent
    - stationary in front of agent
    - pass agent
    - overtake agent
    - moving left of agent
    - moving right of agent
    - stationary left of agent
    - stationary right of agent
    - wait ped cross
    - accelerate
    - decelerate
    - stop
    - lane change
    - u turn
    - reverse
    - left turn
    - right turn
  - stationary left of agent
    - pass agent
    - overtake agent
    - moving left of agent
    - moving right of agent
    - stationary right of agent
    - lead agent
    - follow agent
    - stationary in front of agent
    - stationary behind agent
    - wait ped cross
    - accelerate
    - decelerate
    - stop
    - lane change
    - u turn
    - reverse
    - left turn
    - right turn
  - stationary right of agent
    - pass agent
    - overtake agent
    - moving right of agent
    - moving left of agent
    - stationary left of agent
    - lead agent
    - follow agent
    - stationary in front of agent
    - stationary behind agent
    - wait ped cross
    - accelerate
    - decelerate
    - stop
    - lane change
    - u turn
    - reverse
    - left turn
    - right turn
  - walk alongside
    - walk opposite
    - overtake agent
    - pass agent
    - run
    - stand
    - cross
    - jaywalk
  - stationary in front of agent
    - lead agent
    - follow agent
    - stationary behind agent
    - pass agent
    - overtake agent
    - moving left of agent
    - moving right of agent
    - stationary left of agent
    - stationary right of agent
    - wait ped cross
    - accelerate
    - decelerate
    - stop
    - lane change
    - u turn
    - reverse
    - left turn
    - right turn

- agent-agent
  - stationary behind agent
    - follow agent
    - lead agent
    - stationary in front of agent
    - pass agent
    - overtake agent
    - moving left of agent
    - moving right of agent
    - stationary left of agent
    - stationary right of agent
    - wait ped cross
    - accelerate
    - decelerate
    - stop
    - lane change
    - u turn
    - reverse
    - left turn
    - right turn
  - pass agent
    - overtake agent
    - moving left of agent
    - moving right of agent
    - stationary left of agent
    - stationary right of agent
    - lead agent
    - follow agent
    - stationary in front of agent
    - stationary behind agent
    - wait ped cross
    - accelerate
    - decelerate
    - stop
    - lane change
    - u turn
    - reverse
    - left turn
    - right turn
    - walk opposite
    - jaywalk
    - walk alongside
    - run
    - stand
  - moving left of agent
    - pass agent
    - overtake agent
    - stationary left of agent
    - moving right of agent
    - stationary right of agent
    - lead agent
    - follow agent
    - stationary in front of agent
    - stationary behind agent
    - wait ped cross
    - accelerate
    - decelerate
    - stop
    - lane change
    - u turn
    - reverse
    - left turn
    - right turn
  - moving right of agent
    - pass agent
    - stationary behind agent
    - stationary right of agent
    - overtake agent
    - stationary in front of agent
    - stop
    - decelerate
    - left turn
    - reverse
    - accelerate
    - u turn
    - follow agent
    - moving left of agent
    - lane change
    - stationary left of agent
    - right turn
    - wait ped cross
    - lead agent

- agent-agent
  - walk opposite
    - walk alongside
    - overtake agent
    - pass agent
    - run
    - stand
    - cross
    - jaywalk
  - lead agent
    - stationary behind agent
    - pass agent
    - stationary right of agent
    - overtake agent
    - stationary in front of agent
    - stop
    - decelerate
    - left turn
    - reverse
    - accelerate
    - moving right of agent
    - u turn
    - follow agent
    - moving left of agent
    - lane change
    - stationary left of agent
    - right turn
    - wait ped cross
  - wait ped cross
    - pass agent
    - overtake agent
    - overtake ego
    - follow agent
    - lead agent
    - follow ego
    - lead ego
    - decelerate
    - left turn
    - right turn
    - stop
    - reverse
    - u turn
  - overtake agent
    - pass agent
    - moving left of agent
    - moving right of agent
    - stationary left of agent
    - stationary right of agent
    - lead agent
    - follow agent
    - stationary in front of agent
    - stationary behind agent
    - wait ped cross
    - accelerate
    - decelerate
    - stop
    - lane change
    - u turn
    - reverse
    - left turn
    - right turn
    - walk alongside
    - walk opposite
    - run
    - stand
    - cross
    - jaywalk

Figure 9: **Scenario catalog.** Agent-Agent scenarios with assigned negative scenarios in red.

Table 3: **Ego-to-agent scenario definition.** Textual description of scenarios where the ego-vehicle interacts with other agents.

lead agent	Ego travels ahead of agent at a similar speed while maintaining a consistent distance.
follow agent	Ego is driving behind agent at a similar speed while maintaining a consistent distance.
pass agent	Ego in the adjacent lane overtakes the stopped agent.
overtake agent	Ego in the adjacent lane and moves ahead of agent while both are in motion.
stationary left of agent	Agent is fully stopped and remains stationary to the left of agent, which is also stationary, such as when waiting at a traffic light or in a parking lot.
stationary in front of agent	Agent is fully stopped and remains stationary in front of agent, such as when waiting at a traffic light, in a parking lot, or any other situation requiring queuing.
wait ped cross	Ego comes to a stop or remains stationary, yielding the right-of-way to agent who is crossing or preparing to cross the road, while maintaining awareness of the agent's movement and ensuring a safe distance until the crossing is complete.

Table 4: **Agent scenario definition.** Textual description of agent scenarios.

stand	Agent (pedestrian) remains stationary in the traffic environment, either waiting at a crossing, observing surroundings, or pausing for other reasons.
walk	Agent (pedestrian) moves at a steady, moderate pace, typically following designated paths or crosswalks.
jaywalk	Agent (pedestrian) crosses the street outside of designated crossing areas or against traffic signals, often requiring heightened awareness of vehicle movements, quick decision-making to avoid conflicts, and potentially creating unpredictable interactions with other agents in the traffic environment.
run	Agent (pedestrian) is running and moves rapidly.
cross	Agent (pedestrian) moves from one side of the road to the other, at a designated crossing point or intersection.
accelerate	Agent is increasing its speed, either gradually or abruptly, to adapt to traffic conditions, maintain flow, or comply with traffic rules and signals.
stop	Agent is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.
reverse	Agent is moving in reverse, either to park, navigate a tight space, or adjust its position.
left turn	Agent is executing a left turn at an intersection or junction.
right turn	Agent is executing a right turn at an intersection or junction.
u turn	Agent is performing a 180-degree turn at an intersection or junction, reversing its direction of travel.
lane change	Agent is transitioning from its current lane to an adjacent lane.
lead ego	Agent travels ahead of ego at a similar speed while maintaining a consistent distance.
follow ego	Agent is driving behind ego at a similar speed while maintaining a consistent distance.
overtake ego	Agent in the adjacent lane moves ahead of ego while both are in motion.
stationary right of ego	Agent is fully stopped and remains stationary to the right of ego, which is also stationary, such as when waiting at a traffic light or in a parking lot.
stationary behind ego	Agent is fully stopped and remains stationary behind ego (which is also stopped), such as when waiting at a traffic light, in a parking lot, or in any other queuing scenario.

Table 5: **Ego scenario definition.** Textual description of ego-vehicle scenarios.

accelerate	Ego is increasing its speed, either gradually or abruptly, to adapt to traffic conditions, maintain flow, or comply with traffic rules and signals.
decelerate	Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.
stop	Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.
left turn	Ego is executing a left turn at an intersection or junction.
right turn	Ego is executing a right turn at an intersection or junction.
lane change	Ego is transitioning from its current lane to an adjacent lane.

Table 6: **Agent-to-agent scenario definition.** Textual description of scenarios where agents interact with other agents.

walk alongside	Agent (pedestrian) and agent (pedestrian) walk side by side at a steady, moderate pace.
walk opposite	Agent (pedestrian) and agent (pedestrian) walk toward each other at a moderate pace, cross paths, and proceed.
lead agent	Agent travels ahead of agent at a similar speed while maintaining a consistent distance.
follow agent	Agent is driving behind agent at a similar speed while maintaining a consistent distance.
pass agent	Agent in the adjacent lane overtakes the stopped agent.
overtake agent	Agent in the adjacent lane and moves ahead of agent while both are in motion.
moving left of agent	Agent is traveling in parallel to the left of agent (e.g., in adjacent lanes or side by side), with one vehicle maintaining a leftward offset relative to the other. This could occur during lane-matched driving on a multi-lane road or synchronized movement from a traffic light.
moving right of agent	Agent is traveling in parallel to the right of agent (e.g., in adjacent lanes or side by side), with one vehicle maintaining a rightward offset relative to the other. This could occur during lane-matched driving on a multi-lane road or synchronized movement from a traffic light.
stationary left of agent	Agent is fully stopped and remains stationary to the left of agent, which is also stationary, such as when waiting at a traffic light or in a parking lot.
stationary right of agent	Agent is fully stopped and remains stationary to the right of agent, which is also stationary, such as when waiting at a traffic light or in a parking lot.
stationary in front of agent	Agent is fully stopped and remains stationary in front of agent, such as when waiting at a traffic light, in a parking lot, or any other situation requiring queuing.
stationary behind agent	Agent is fully stopped and remains stationary behind agent, such as when waiting at a traffic light, in a parking lot, or any other situation requiring queuing.
wait ped cross	Agent comes to a stop or remains stationary, yielding the right-of-way to a agent who is crossing or preparing to cross the road, while maintaining awareness of the agent's movement and ensuring a safe distance until the crossing is complete.

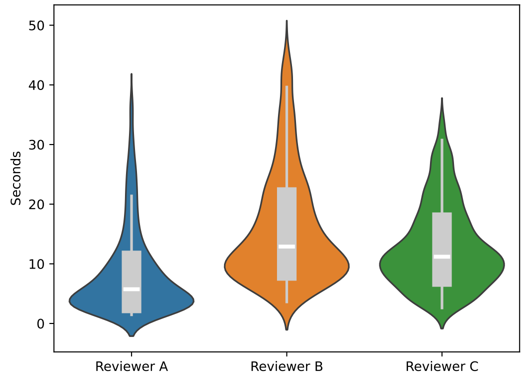


Figure 10: **Validation time.** Comparison of the validation duration for each reviewer.

A.3 Scenario verification

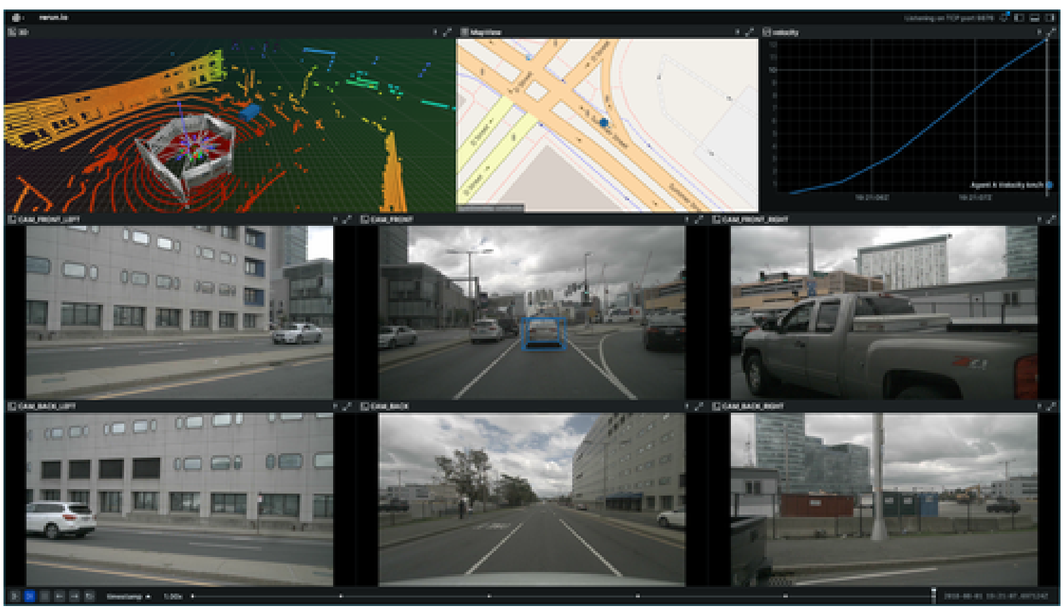
Verification tool. Fast and simple verification requires a clean and versatile visualization interface paired with a simple input mask. In Fig. 11, we show the verification command line tool (a) of STS-Bench, that can be used via mouse or keyboard only. The visualization interface (b) uses Rerun [50] and shows various available input modalities, such as LiDAR point clouds, agent trajectories on a map, velocities, and multi-view camera images for consecutive time frames. The presented validation tool enables fast inspection of mined scenarios in different modalities and effortless acceptance or rejection of samples.

Verification insights. To demonstrate the efficiency of our verification, we provide a detailed time analysis in Fig. 10. We observe an average of 8.3 seconds, 16.0 seconds, and 12.9 seconds for reviewers A, B, and C, respectively. While some scenes require more attention, most of the scenarios are obvious and can therefore be confirmed quickly.

In the following, we show typical verification mistakes and disagreements between reviewers leading to either accepted or rejected scenarios. These verification differences are solved by majority voting. Fig. 12 shows a highly occluded person (blue), which has been considered too difficult by one of the three reviewers. However, although only a small part of the person is visible, this sample can serve as a difficult scenario for the benchmark. The next two scenes in Fig. 13 and Fig. 14 illustrate scenarios that the reviewers have rejected because they do not follow the scenario definition exactly. In Fig. 13, the ego-vehicle *passing* the agent (green) is not in the adjacent line. Fig. 14 shows a similar issue, where the agent (green) *following* the ego-vehicle has another agent in between, which is not considered following the ego-vehicle in our scenario definition. For *jaywalking*, the definition states that a pedestrian crosses the street outside designated crossing areas. However, not all designated areas are properly annotated in map data, as shown in Fig. 15. Construction workers operating in a safe construction space (Fig. 15a) or pedestrians crossing a driveway between two sidewalks (Fig. 15b) are not considered jaywalking.

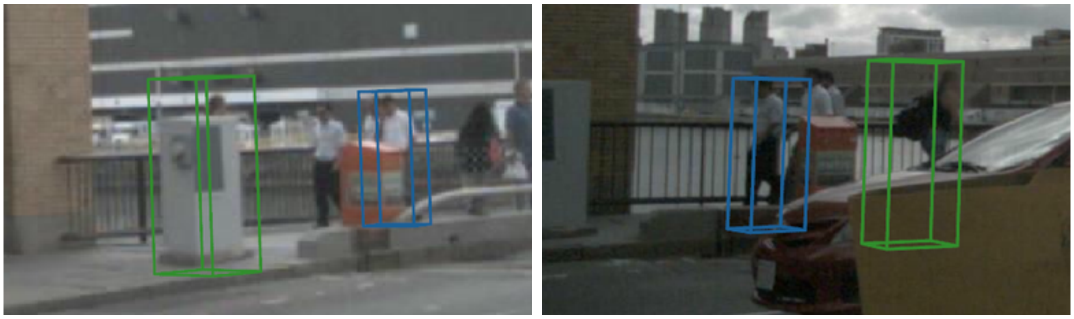
Disagreements can also occur due to inaccurate visualizations of projected 3D bounding boxes onto the camera image. In Fig. 16, we illustrate an example where bounding boxes indicate two neighboring vehicles (green, blue) in the camera image (a). However, a closer look at the LiDAR scan (b) reveals another agent between them. Consequently, the ambiguous verification of such samples could either stem from the inaccurate projection or a misunderstanding of the scenario description. Finally, subjective perception can also lead to ambiguous verification results. Fig. 17 shows a lane change scenario hardly recognizable from the trajectory in the map view (b). However, the vehicle (blue) is heavily occluded, and lane markings are not visible in the camera image (a). Thus, such hardly recognizable scenarios have been rejected in the manual verification stage.

(a) Verification GUI.



(b) Verification rerun.

Figure 11: Verification tool for fast and simple verification (a) of mined traffic scenarios by inspecting (b) recordings from various available modalities.



(a) First frame (CAM_FRONT_LEFT). (b) Fourth frame (CAM_FRONT_LEFT).

Figure 12: **Agent walk opposite.** Accepted scenario sample after review, since the occlusion level has been rated difficult but reasonable. The bounding box is not visible for the tested models.

(a) First frame (CAM_FRONT_RIGHT). (b) Last frame (CAM_BACK).

Figure 13: **Ego-vehicle passes agent.** The scenario sample was rejected after review since the ego-vehicle is not in the adjacent lane of the refereed agent. The bounding box is not visible for the tested models.

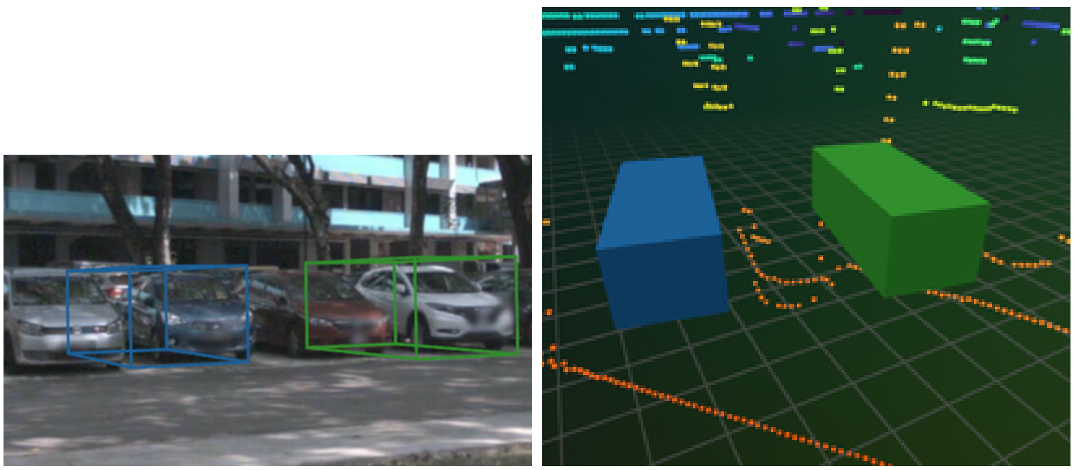
(a) First frame (CAM_BACK). (b) Last frame (CAM_BACK).

Figure 14: **Agent following ego-vehicle.** Rejected scenario sample after review since there is another agent between the ego-vehicle and the refereed agent. The bounding box is not visible for the tested models.

(a) Construction worker in a safe space.

(b) Pedestrian on sidewalk.

Figure 15: **Agent jaywalking.** Rejected scenario sample after review since there is no jaywalking. The bounding box is not visible for the tested models.



(a) First frame (CAM_FRONT_LEFT).

(b) 3D LiDAR scene.

Figure 16: **Agent right of agent.** The scenario sample was rejected after review since there is another agent between the two involved agents. The bounding box is not visible for the tested models.

(a) First frame (CAM_BACK_RIGHT).

(b) Trajectory on map.

Figure 17: **Agent lane change.** The scenario sample was rejected after review since no clear lane change was visible. The bounding box is not visible for the tested models.

Table 7: **End-to-end driving models.** Driving expert VLMs for end-to-end driving, *i.e.*, open-loop planning (OLP), closed-loop driving (CLD), and control signal prediction (CSP). Additional proxy tasks are scene generation (SG), dense captioning (CAP), and counterfactual reasoning (CR). Rows marked in gray provide explicit reasoning, code, and pre-trained model weights. Code marked with "~" does not include evaluation scripts or configurations for NuScenes.

Method	Evaluation Dataset	Environment	Task	Explicit reasoning	Model Weights	Code
ORION [18]	NuScenes, CARLA [17]	Real, Sim	CLD, OLP	✓	✓	~
DriveMM [22]	NuScenes	Real	VQA	✓	✓	✓
RDA-Driver [23]	NuScenes	Real	OLP	✓	✗	✗
EMMA [24]	WOD, NuScenes	Real	OLP	✓	✗	✗
Senna-VLM [29]	NuScenes, DriveX [29]	Real	CSP	✓	✓	✓
GPVL [34]	NuScenes	Real	OLP	✓	✗	✓
VLP [47]	NuScenes	Real	OLP	✗	✗	✗
LMDrive [52]	CARLA [17]	Sim	CLD	✗	✓	✓
DriveVLM-Dual [53]	NuScenes	Real	OLP	✓	✗	✗
InsightDrive [54]	NuScenes	Real	OLP	✓	✗	✗
OmniDrive [60]	NuScenes	Real	OLP, CR	✓	✓	✓
BEVDriver [63]	CARLA [17]	Sim	CLD	✗	✗	✗
Sce2DriveX [71]	NuScenes	Real	OLP, CSP	✓	✗	✗
HERMES [74]	NuScenes	Real	SG, OLP	✓	✗	✗
OpenDriveVLA [75]	NuScenes	Real	OLP, CAP	✓	✗	✗

A.4 Multiple-choice question generation

The evaluation on our benchmark requires the generation of input prompts that contain the task description and a question. In addition to the preamble (setting the context) and scenario descriptions (for all answer choices), we ask the models one of the following questions, depending on the scenario category:

- **Ego:** Which of the following options best describes ego driving maneuver?
- **Agent:** Which of the following options best describes the driving behavior of the <reference to agent>?
- **Ego-to-agent:** Which of the following options best describes the ego driving behavior with respect to the <reference to agent>?
- **Agent-to-agent:** Which of the following options best describes <reference to agent 1> maneuver with respect to the <reference to agent 2>?

Notice that this example is only valid for DriveMM [22] and requires adaptation for other models in order to get better performance. For instance, the fine-tuning of Senna-VLM [29] always refers in a first-person manner to the ego-vehicle, which requires altering questions respectively. Detailed prompting examples for all evaluated models are provided in Sec.E.

B Driving expert baselines

In this section, we discuss the choice of driving expert models from recent publications listed in Table 7. To be considered for evaluation, the model must have explicit reasoning capabilities in addition to its primary task. Furthermore, the code and model weights must be publicly available to guarantee fair comparison. Therefore, we opted for DriveMM [22], Senna-VLM [29], and OmniDrive [60].

DriveMM. Not particularly developed for the end-to-end driving task, DriveMM [22] performs joint training on multiple driving datasets. The model architecture is inspired by LLaVa [38] and comprises a SigLIP [70] vision encoder and Llama-3.1 [1] as LLM. In the first step, the model gets pre-trained on a multi-modal dataset that includes text-image pairs for images and videos. For the final fine-tuning, the model learns from single images, multi-view images, and videos, leveraging various driving datasets, *i.e.*, DriveLM [53], LingoQA [44], NuInstruct [16], OmniDrive [60], MAPLM [7], and CODA-LM [35]. Hence, the model learns to process various input modalities from single images to multi-view videos.

Table 8: **Ego scenario evaluation.** Scenario-level performance comparison of LLMs, off-the-shelf VLMs, and driving expert VLMs for ego scenarios. Accuracy grouped by scenario categories: Left Turn (**LT**), Accelerate (**Acc**), Right Turn (**RT**), Lane Change (**LC**), Decelerate (**Dec**), Stop and Average (**Avg**). The best results are highlighted in bold.

	LT	**Acc**	**RT**	**LC**	**Dec**	**Stop**	**Avg**
Llama 3.2 [1]	9.1	24.3	0.0	18.8	29.4	18.2	16.6
DeepSeek V3 [12]	27.3	86.5	10.0	12.5	41.2	9.1	31.1
GPT-4o [46]	**100.0**	**91.9**	50.0	**68.8**	52.9	18.2	**63.6**
InternVL 2.5 1B [9]	0.0	29.7	20.0	37.5	23.5	9.1	20.0
Qwen2.5-VL 7B [5]	36.4	2.7	**70.0**	6.3	52.9	45.5	35.6
InternVL 2.5 8B [9]	18.2	21.6	50.0	6.3	**70.6**	**63.6**	38.4
Senna-VLM [29]	18.2	10.8	0.0	6.3	17.7	0.0	8.8
OmniDrive [60]	25.0	25.0	23.8	35.3	27.8	25.0	24.4
DriveMM [22]	63.6	51.4	**70.0**	18.8	11.8	45.5	43.5

OmniDrive. This model is designed to provide driving decisions and waypoints for planning. It encodes multi-view images into BEV features leveraging the StreamPETR [59] encoder and maps this representation into language space, incorporating a Q-Former [33]. The training contains a series of objectives, including 3D grounding, open-loop planning, counterfactual reasoning, and scene understanding. Annotations for training on the NuScenes [6] dataset are automatically generated leveraging ground truth information and VLMs.

Senna. This end-to-end driving model consists of two modules, Senna-VLM and SennaE2E. The idea is to predict high-level planning decisions from multi-view video inputs with Senna-VLM and encode them into meta-action features, enhancing the final planning trajectories of SennaE2E, which is a state-of-the-art end-to-end driving planner [28]. The images are encoded with CLIP [49] and compressed in order to reduce the number of input tokens. The model is trained in multiple steps, starting with pre-training on multiple sources, including the instruction following data from LlaVa [38]. Two fine-tuning steps follow on private data with automatically generated VLM annotations, leveraging ground truth information. First, the model learns to describe the scene, detect traffic signals, identify vulnerable road users, and so on. Second, the model learns to predict meta-actions such as accelerate, decelerate, left, right, *etc*.

C Detailed results & analysis

We provide detailed evaluations for ego, agent, ego-to-agent, and agent-to-agent scenarios in Table 8, 9, 10, and 11, respectively. Therefore, we list the accuracy for each scenario and the weighted average for the corresponding category. In addition to evaluating the performance of the individual models, we also gain the following insights:

Ego scenarios. Large language models (LLMs) perform exceptionally well for ego scenarios, shown in Table 8, as they have the ground truth trajectory of the ego-vehicle at their disposal. Especially GPT-4o [46] with advanced reasoning can successfully derive scenarios from given trajectories. We illustrate the model's reasoning after correctly identifying a *left turn* in Fig. 19. An exception is the differentiation between the closely related scenarios *decelerate* and *stop*. Since the vehicle may not be fully stationary in the trajectory data, such as when moving at less than 1 km/h, the language model typically interprets the situation as deceleration. In contrast, powerful vision-language models (VLMs) often recognize it as a potential stop, based on contextual cues like a vehicle ahead, a red traffic light, or a stop sign. However, end-to-end driving models fail to perform well across different scenarios. Only DriveMM [22] provides reasonable results for more obvious scenarios such as *left turn* or *right turn*.

Agent scenarios. Table 9 shows the detailed evaluation of various models on agent scenarios. These scenarios that involve a single agent are more difficult for LLMs. A simple reason is that some of these scenarios, *i.e.*, *jaywalking*, *crossing*, require road markings to be detected correctly. Surprisingly,

Table 9: **Agent scenario evaluation.** Scenario-level performance comparison of LLMs, off-the-shelf VLMs, and driving expert VLMs for ego-to-agent scenarios. Accuracy grouped by scenario categories: Jaywalk (**JW**), Left Turn (**LT**), Walk (**W**), Stand (**S**), Cross (**C**), Accelerate (**Acc**), Right Turn (**RT**), Lane Change (**LC**), Overtake Ego (**OE**), Follow Ego (**FE**), Stop, Run, Lead Ego (**LE**), Reverse (**Rev**), U Turn (**UT**), Stationary Right Of Ego (**SRoE**), Stationary Behind Ego (**SBE**) and Average (**Avg**). The best results are highlighted in bold.

	JW	LT	W	S	C	Acc	RT	LC	OE	FE	Stop	Run	LE	Rev	UT	SRoE	SBE	Avg
Llama 3.2 [1]	16.0	16.7	20.8	9.1	10.4	32.0	15.4	29.2	0.0	20.5	10.3	20.0	35.3	0.0	0.0	0.0	0.0	13.8
DeepSeek V3 [12]	28.0	45.8	33.3	13.6	41.7	20.0	30.8	12.5	**83.3**	89.7	20.7	0.0	**94.1**	0.0	0.0	**100.0**	**100.0**	42.0
GPT-4o [46]	36.0	**62.5**	25.0	13.6	33.3	52.0	38.5	16.7	33.3	**94.9**	6.9	**80.0**	82.4	0.0	0.0	**100.0**	**100.0**	45.6
InternVL 2.5 1B [9]	16.0	29.2	35.4	0.0	70.8	22.0	26.9	16.7	66.7	15.4	17.2	20.0	11.7	0.0	0.0	0.0	0.0	20.5
Qwen2.5-VL 7B [5]	4.0	29.2	68.8	72.7	79.2	2.0	42.3	20.8	33.3	61.5	10.3	**80.0**	17.7	0.0	0.0	50.0	50.0	36.6
InternVL 2.5 8B [9]	8.0	16.7	**85.4**	**75.0**	**89.6**	28.0	**65.4**	33.3	33.3	64.1	20.7	40.0	47.1	**100.0**	0.0	**100.0**	**100.0**	**53.3**
Senna-VLM [29]	**32.0**	0.0	58.3	31.8	35.4	30.0	30.8	12.5	33.3	61.5	6.9	20.0	41.2	0.0	0.0	50.0	0.0	26.1
OmniDrive [60]	26.3	44.8	32.4	14.5	50.6	28.2	45.9	3.2	66.7	20.6	27.3	28.6	15.2	0.0	0.0	0.0	0.0	23.8
DriveMM [22]	0.0	16.7	75.0	22.7	50.0	**82.0**	30.8	8.3	16.7	79.5	**37.9**	0.0	76.5	0.0	0.0	50.0	**100.0**	38.0

Table 10: **Ego-to-agent scenario evaluation.** Scenario-level performance comparison of LLMs, off-the-shelf VLMs, and driving expert VLMs for ego-to-agent scenarios. Accuracy grouped by scenario categories: Lead Agent (**LA**), Pass Agent (**PA**), Wait Pedestrian Cross (**WPC**), Follow Agent (**FA**), Overtake Agent (**OA**), Stationary Left Of Agent (**SLoA**), Stationary In Front Of Agent (**SiFoA**) and Average (**Avg**). The best results are highlighted in bold.

	LA	**PA**	**WPC**	**FA**	**OA**	**SLoA**	**SiFoA**	**Avg**
Llama 3.2 [1]	27.5	18.2	57.1	35.3	0.0	0.0	0.0	19.7
DeepSeek V3 [12]	50.0	**72.7**	57.1	70.6	**100.0**	50.0	**100.0**	71.5
GPT-4o [46]	**77.5**	25.0	57.1	70.6	**100.0**	**100.0**	**100.0**	**75.8**
InternVL 2.5 1B [9]	17.5	68.2	78.6	52.9	0.0	0.0	50.0	38.2
Qwen2.5-VL 7B [5]	10.0	0.0	**100.0**	76.5	0.0	0.0	0.0	26.6
InternVL 2.5 8B [9]	32.5	9.1	**100.0**	**94.1**	0.0	0.0	50.0	40.8
Senna-VLM [29]	27.5	27.3	85.7	70.6	**100.0**	0.0	0.0	44.4
OmniDrive [60]	32.4	44.1	80.0	19.4	**100.0**	0.0	25.0	43.0
DriveMM [22]	**77.5**	29.6	92.9	88.2	0.0	0.0	50.0	48.3

GPT-4o [46] performs above random guessing. However, this outcome is less indicative of true scenario understanding and more a reflection of the model's strong general reasoning capabilities. We illustrate this in Fig. 18, where GPT-4o [46] arrives at the correct answer through a process of elimination. The best overall performance for this category reaches InternVL 2.5 8B [9]. The VLM has difficulties with challenging scenarios like *jaywalking* and *lane change*, requiring in-depth visual understanding, but also with scenarios that involve the ego vehicle, such as *overtaking ego*, *lead ego*, and *follow ego*. The relatively bad performance of driving expert models highlights the need to enhance spatio-temporal understanding for these models.

Ego-agent scenarios. The detailed results for all ego-to-agent scenarios are listed in Table 10. We observe that LLMs perform very well at this task. However, we also notice a respectable performance of driving expert models. This is unsurprising since these scenarios are also part of existing benchmarks and datasets. The objective of driving experts mostly includes reasoning about agents in their close vicinity. The most challenging cases are *pass agent*, and both scenarios of identifying *stationary* objects in a certain position, *i.e.*, *left* and *right* of the ego-vehicle. The performance on these relatively unknown but simple scenarios highlights a strong bias toward previously seen scenarios and underscores the limited zero-shot generalization capabilities of these models.

Agent-to-agent scenarios. The interactions between two agents not involving the ego-vehicle are the most difficult to reason about for the assessed models. In Table 11, InternVL 2.5 8B [9] reaches the best average score across all scenarios. We can observe that OmniDrive [60] is surprisingly good

Table 11: **Agent-to-agent scenario evaluation.** Scenario-level performance comparison of LLMs, off-the-shelf VLMs, and driving expert VLMs for Agent-to-agent scenarios. Accuracy grouped by scenario categories: Follow Agent (**FA**), Stationary Left Of Agent (**SLoA**), Stationary Right Of Agent (**SRoA**), Walk Alongside (**WA**), Stationary In Front Of Agent (**SiFoA**), Walk Opposite (**WO**), Lead Agent (**LA**), Wait Pedestrian Cross (**WPC**), Overtake Agent (**OA**), Stationary Behind Agent (**SBA**), Pass Agent (**PA**), Moving Left Of Agent (**MLoA**), Moving Right Of Agent (**MRoA**) and Average (**Avg**). The best results are highlighted in bold.

	FA	**SLoA**	**SRoA**	**WA**	**SiFoA**	**WO**	**LA**	**WPC**	**OA**	**SBA**	**PA**	**MLoA**	**MRoA**	**Avg**
Llama 3.2 [1]	43.2	5.9	16.7	25.0	12.5	32.5	37.8	20.0	25.0	11.1	20.9	33.3	0.0	21.8
DeepSeek V3 [12]	52.3	5.9	8.3	65.0	6.3	**95.0**	26.7	96.0	25.0	5.6	55.8	33.3	**100.0**	44.2
GPT-4o [46]	**65.9**	5.9	5.6	**75.0**	6.3	**95.0**	**71.1**	84.0	41.7	0.0	30.2	33.3	50.0	43.4
InternVL 2.5 1B [9]	6.8	2.9	2.8	17.5	**25.0**	27.5	24.4	40.0	**50.0**	16.7	**58.1**	66.7	50.0	29.9
Qwen2.5-VL 7B [5]	63.6	23.5	36.1	55.0	**25.0**	45.0	20.0	84.0	25.0	38.9	20.9	0.0	50.0	37.5
InternVL 2.5 8B [9]	43.2	**61.8**	**63.9**	72.5	0.0	72.5	26.7	**100.0**	0.0	**50.0**	32.6	**100.0**	50.0	**51.8**
Senna-VLM [29]	63.6	23.5	25.0	30.0	**25.0**	27.5	28.9	68.0	25.0	16.7	25.6	0.0	50.0	31.5
OmniDrive [60]	54.7	12.7	13.5	1.8	9.4	4.5	44.3	57.1	**52.9**	32.6	56.5	0.0	0.0	26.2
DriveMM [22]	34.1	35.3	55.6	32.5	6.3	17.5	31.1	24.0	8.3	27.8	11.6	33.3	50.0	28.3

at scenarios like *overtaking* and *passing agent*. However, the model does not recognize pedestrian scenarios, *i.e. walk opposite* and *walk alongside*, and is not able to identify positions of objects, *i.e.*, *left* and *right* for both motion states *stationary* and *moving*. In contrast, InternVL 2.5 8B [9] is exceptionally good at scenarios that can be identified by visual cues, but does not work well for scenarios spanned across different camera viewpoints, such as *overtaking* or *passing agent*.

Multiple-choice letter distribution. To mitigate potential biases arising from uneven ground truth letter distributions, our experimental design employs a uniform distribution across all multiple-choice options. This critical aspect, visualized in Fig. 20, addresses the known tendency of some LLMs to favor specific letter choices [72]. While methods such as GPT-4o [46], InternVL 2.5 8B [9], and DriveMM [22] appear to be free of this bias, our analysis reveals that Llama 3.2 [1] exhibits a preference for alphabetically earlier letters, a pattern similarly observed in the Senna-VLM [29] expert model. Furthermore, Senna-VLM [29] demonstrates a tendency to disregard prompt instructions, such as the restriction to single-letter outputs, and occasionally includes irrelevant information (most probably from its training data), as illustrated in Fig. 37. The overall output vocabulary of Senna-VLM is depicted in the word cloud in Fig. 21c.

System Prompt:

```
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers
performed by various agents in diverse driving scenarios.
```

User Prompt:

```
An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction
vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. The traffic environment is captured using a
sophisticated suite of sensors mounted on an ego vehicle, which is a standard passenger car.
The sensor suite includes a Light Detection and Ranging (LiDAR) sensor labeled LIDAR_TOP, mounted on the roof of the car.
This LiDAR sensor provides high-precision 3D spatial data about the surrounding environment. LiDAR data is crucial for
precise spatial positioning and size, which helps in differentiating vehicle types and detecting movement. For each agent,
the LiDAR data includes the frame number, the center position of the agent (x, y, z) in meters relative to the LiDAR, and a
heading angle in degrees relative to the LiDAR, and size dimensions such as width, length, and height in meters.
In addition to the LiDAR, six cameras are strategically positioned around the vehicle to provide a 360-degree field of view.
Camera data can aid in visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned
directly in front of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented
at a 45-degree angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a
135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the
left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally,
CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. For each
agent, the camera data includes the frame number, the center of the agent given with pixel location (x, y) in the image
frame and in which camera is agent visible (e.g. CAM_FRONT or CAM_BACK).
The system also integrates a GPS sensor, which provides the ego vehicle's precise global position. The GPS data includes
the frame number, coordinates (x, y) in meters within a global coordinate system and the vehicle's orientation in radians
relative to the global frame.
Together, this comprehensive sensor suite enables detailed monitoring and analysis of the dynamic behaviors of all traffic
agents. Your task is to leverage this data to identify and interpret the temporal actions and maneuvers of each agent
within the traffic environment.
The following are driving maneuvers and actions along with their respective descriptions:
Agent 1 crossing street: Agent 1 (pedestrian) moves from one side of the road to the other, at a designated crossing point
or intersection.
Agent 1 jaywalking: Agent 1 (pedestrian) crosses the street outside of designated crossing areas or against traffic signals,
often requiring heightened awareness of vehicle movements, quick decision-making to avoid conflicts, and potentially
creating unpredictable interactions with other agents in the traffic environment.
Agent 1 running: Agent 1 (pedestrian) is running and moves rapidly.
Agent 1 decelerating: Agent 1 is reducing its speed, either gradually or abruptly, in response to traffic conditions,
obstacles, or to comply with traffic rules, without coming to a complete stop.
Agent 1 stopping: Agent 1 is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles,
or to comply with traffic rules and comes to a complete stop.
Your input consists of sequential data, captured over 6 frames and 6 seconds with the described sensor suite.
Ego:
Frame number: 0
x: 0.00
y: 0.00
rotation: 0.00
...
Frame number: 5
x: 7.95
y: 0.33
rotation: 0.03
Agent 1:
Class: pedestrian
Frame number: 0
LiDAR x: 6.45
LiDAR y: -22.25
LiDAR rotation: 2.93
CAM x: 580.66
CAM y: 509.22
CAM: CAM_BACK
...
Frame number: 5
LiDAR x: 4.63
LiDAR y: -29.92
LiDAR rotation: 2.90
CAM x: 696.33
CAM y: 513.36
CAM: CAM_BACK
Which of the following options best describes agent 1 driving maneuver?
Options:
A. Agent 1 crossing street
B. Agent 1 jaywalking
C. Agent 1 running
D. Agent 1 decelerating
E. Agent 1 stopping
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.
```

Answer:

```
GPT-4o [46]: B                                                                                                 Correct: B
```

User Prompt:

```
Can you elaborate your decision?
```

Answer:

```
GPT-4o [46]: Agent 1 is moving consistently and progressively away from the ego vehicle without any indication of stopping,
running rapidly, or interacting with specific crossing points. The movement doesn't suggest crossing at a designated area,
so the best interpretation given the information is that Agent 1 is likely crossing outside designated areas, which aligns
with jaywalking behavior.
```

Figure 18: GPT-4o [46] reasoning for a jaywalking scenario.

System Prompt:

```
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers
performed by various agents in diverse driving scenarios.
```

User Prompt:

```
An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction
vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles.  The traffic environment is captured using a
sophisticated suite of sensors mounted on an ego vehicle, which is a standard passenger car.
The sensor suite includes a Light Detection and Ranging (LiDAR) sensor labeled LIDAR_TOP, mounted on the roof of the car.
This LiDAR sensor provides high-precision 3D spatial data about the surrounding environment.  LiDAR data is crucial for
precise spatial positioning and size, which helps in differentiating vehicle types and detecting movement.  For each agent,
the LiDAR data includes the frame number, the center position of the agent (x, y, z) in meters relative to the LiDAR, and a
heading angle in degrees relative to the LiDAR, and size dimensions such as width, length, and height in meters.
In addition to the LiDAR, six cameras are strategically positioned around the vehicle to provide a 360-degree field of view.
Camera data can aid in visual confirmation and potentially understanding intent.  The first camera, CAM_FRONT, is positioned
directly in front of the LiDAR sensor and faces forward.  To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented
at a 45-degree angle relative to the front-facing camera.  On the right side of the car, CAM_BACK_RIGHT is positioned at a
135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the
left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera.  Finally,
CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. For each
agent, the camera data includes the frame number, the center of the agent given with pixel location (x, y) in the image
frame and in which camera is agent visible (e.g.  CAM_FRONT or CAM_BACK).
The system also integrates a GPS sensor, which provides the ego vehicle's precise global position.  The GPS data includes
the frame number, coordinates (x, y) in meters within a global coordinate system and the vehicle's orientation in radians
relative to the global frame.
Together, this comprehensive sensor suite enables detailed monitoring and analysis of the dynamic behaviors of all traffic
agents.  Your task is to leverage this data to identify and interpret the temporal actions and maneuvers of each agent
within the traffic environment.
The following are driving maneuvers and actions along with their respective descriptions:
Ego stopping:  Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to
comply with traffic rules and comes to a complete stop.
Ego turning left:  Ego is executing a left turn at an intersection or junction.
Ego performing u-turn:  Ego is performing a 180-degree turn at an intersection or junction, reversing its direction of
travel.
Ego turning right:  Ego is executing a right turn at an intersection or junction.
Ego decelerating:  Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or
to comply with traffic rules, without coming to a complete stop.
Your input consists of sequential data, captured over 6 frames and 6 seconds with the described sensor suite.
Ego:
Frame number:  0
x:  0.00
y:  0.00
rotation:  -0.00
.
.
.
Frame number:  5
x:  7.75
y:  3.50
rotation:  0.82

Which of the following options best describes ego driving maneuver?
Options:
A. Ego stopping
B. Ego turning left
C. Ego performing u-turn
D. Ego turning right
E. Ego decelerating
Please answer only with the letter of an option from the multiple choice list, e.g.  A or B or C or D, and nothing else.
```

Answer:

```
GPT-4o [46]:  B                                                                                Correct:  B
```

User Prompt:

```
Can you elaborate your decision?
```

Answer:

```
GPT-4o [46]:  The ego vehicle's motion data shows increasing x and y coordinates along with a gradually increasing rotation
angle from frames 0 to 5.  The trajectory and rotation suggest the vehicle is moving in an arc to the left.  This pattern
aligns with the maneuver described as "Ego turning left," hence the choice B.
```

Figure 19: GPT-4o [46] reasoning for an ego left turning scenario.

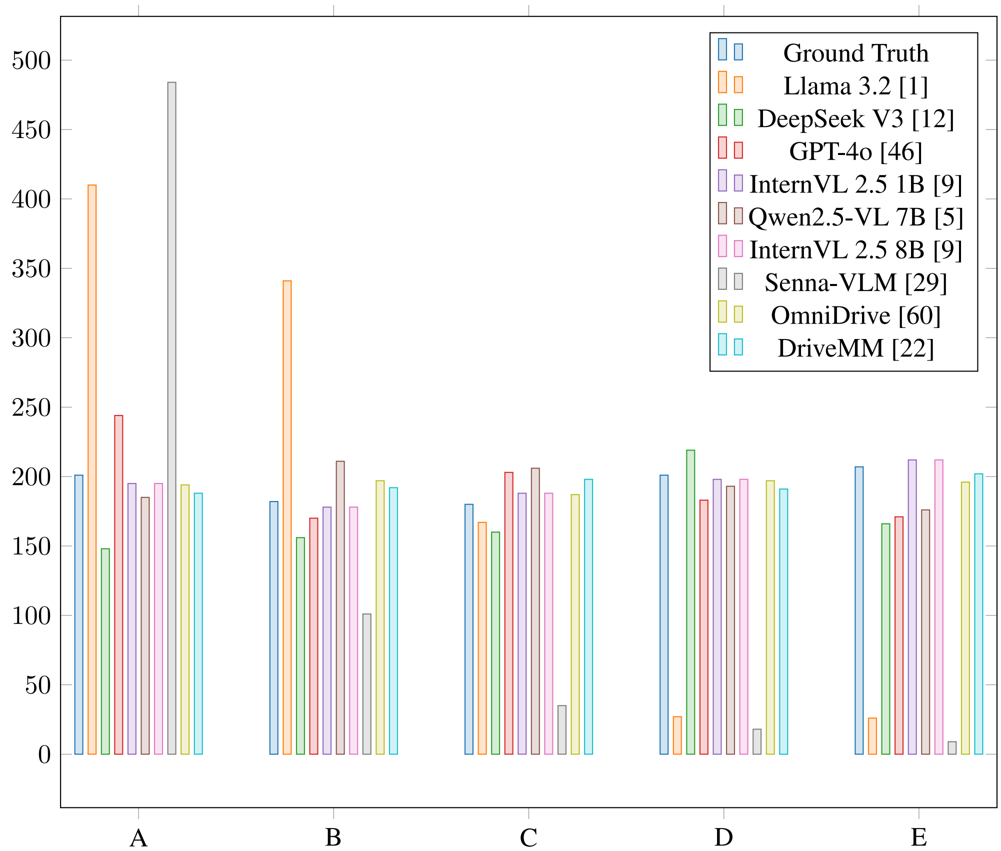


Figure 20: Distribution of individual letters that were the correct multiple-choice option, comparing ground truth data with the predictions of each evaluated method.

(a) Ground truth

(b) Llama 3.2 [1]

(c) Senna-VLM [29]

Figure 21: Distribution of individual letters that were the correct multiple-choice option (a), and the outputs generated by the Llama 3.2 [1] (b) and Senna-VLM [29] (c) models.

Table 12: OmniDrive [60] ablation by querying the model using n^{th} frame and with or without the Chain of Thought (CoT).

	Ego	Ego-to-Agent	Agent	Agent-to-Agent	Average
1st w/ CoT	24.40	42.97	23.78	26.15	29.33
3rd w/ CoT	20.89	**46.92**	**32.93**	**27.24**	**32.00**
6th w/ CoT	21.70	43.65	23.60	26.72	28.92
1st w/o CoT	**29.91**	35.87	21.76	25.39	28.23

Table 13: Expert model evaluation where complete ground truth information was included in the prompt, mirroring the evaluation of LLM-only models.

	Ego	Ego-to-Agent	Agent	Agent-to-Agent	Average
Senna-VLM [29]	3.46	0.00	4.35	4.81	3.16
OmniDrive [60]	27.02	34.61	22.75	29.27	28.41
DriveMM [22]	**53.11**	**71.41**	**39.54**	**29.91**	**48.49**

D Additional experiments

Ablation of the query frame. The default model setting is to provide the initial frame of the sequence as the query frame. We evaluate how the performance is affected by providing a later query frame as the initial input. Specifically, we evaluate OmniDrive [60] by querying with the first (default), middle, and last frames in each scenario. Table 12 shows that querying with the middle frame achieves the highest performance, presumably because it integrates more comprehensive temporal information than the alternatives. Conversely, using the last frame as the query leads to a minor decrease in performance, which we hypothesize is due to limitations in OmniDrive's capacity for long-range temporal dependency modeling.

OmniDrive [60] chain-of-though (CoT). OmniDrive leverages the chain-of-thought (CoT) [61] capabilities of vision-language models (VLMs) to infer trajectories in a sequential, step-by-step manner, starting from scene descriptions, 3D grounding, and other relevant contextual information. In our evaluation, we mirrored this approach by prompting the model at the conclusion of its internal CoT process. This design choice aims to ensure the model encapsulates crucial spatio-temporal information necessary for answering our queries. To ablate the impact of CoT, we conducted an evaluation of OmniDrive while disabling the CoT mechanism (Table 12). As anticipated, the model's performance experienced a slight degradation without CoT, underscoring its positive contribution to overall performance.

Number of multiple-choice options. To allow for varying difficulty levels, our benchmark enables scaling the number of multiple-choice options. Consistent with expectations, Fig. 22 reveals a decrease in DriveMM [22] model accuracy as the number of options grows. This increase in options also leads to higher model variance, suggesting greater difficulty in selecting the correct answer. We find that 5 multiple-choice options represents an optimal balance, preventing the benchmark from being trivial or excessively difficult.

Expert model LLMs with ground truth information. To further evaluate the expert models, we provided them with complete ground truth information in the prompt, mirroring the evaluation of LLM-only models. As shown in Table 13, this significantly degraded the performance of Senna-VLM [29], suggesting that its training compromises the LLM's generalization capabilities, consistent with our findings in Fig. 21c. The performance of OmniDrive [60] remained consistent, indicating an inability to effectively utilize information outside its training data. Notably, DriveMM [22]'s performance increased by nearly 9%, surpassing DeepSeek V3 [12]. Given that DriveMM is trained on six diverse datasets, this highlights the importance of diverse training data in preserving the generalization abilities of LLMs.

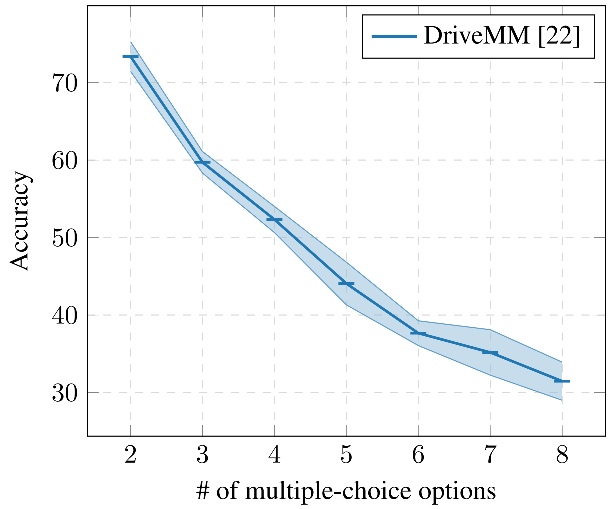


Figure 22: Mean and variance of the DriveMMaccuracy across 5 independent evaluations, shown for increasing numbers of multiple-choice options (*e.g.*, two: A-B; three: A-B-C). In each evaluation, the order of the multiple-choice options was randomly permuted.

E Prompt examples

To demonstrate the required prompt adaptations, we include examples for all model types and scenario types (Ego, agent, ego-to-agent, and agent-to-agent) in the following. In particular, Fig. 23–26 show the LLM prompts. For the representative off-the-shelf VLMs, Fig. 27–30 provide prompt examples for InternVL 2.5 [9], while Fig. 31–34 provide prompt examples for Qwen2.5-VL 7B [4]. For the driving expert models, we provide exemplary prompts for each scenario type for Senna-VLM [29] in Fig. 35–38, OmniDrive [60] in Fig. 39–42, and DriveMM [22] in Fig. 43–46.

```
System Prompt:
```

```
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers
performed by various agents in diverse driving scenarios.
```

```
User Prompt:
```

```
An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction
vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles.  The traffic environment is captured using a
sophisticated suite of sensors mounted on an ego vehicle, which is a standard passenger car.
The sensor suite includes a Light Detection and Ranging (LiDAR) sensor labeled LIDAR_TOP, mounted on the roof of the car.
This LiDAR sensor provides high-precision 3D spatial data about the surrounding environment.  LiDAR data is crucial for
precise spatial positioning and size, which helps in differentiating vehicle types and detecting movement.  For each agent,
the LiDAR data includes the frame number, the center position of the agent (x, y, z) in meters relative to the LiDAR, and a
heading angle in degrees relative to the LiDAR, and size dimensions such as width, length, and height in meters.
In addition to the LiDAR, six cameras are strategically positioned around the vehicle to provide a 360-degree field of view.
Camera data can aid in visual confirmation and potentially understanding intent.  The first camera, CAM_FRONT, is positioned
directly in front of the LiDAR sensor and faces forward.  To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented
at a 45-degree angle relative to the front-facing camera.  On the right side of the car, CAM_BACK_RIGHT is positioned at a
135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the
left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera.  Finally,
CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. For each
agent, the camera data includes the frame number, the center of the agent given with pixel location (x, y) in the image
frame and in which camera is agent visible (e.g.  CAM_FRONT or CAM_BACK).
The system also integrates a GPS sensor, which provides the ego vehicle's precise global position.  The GPS data includes
the frame number, coordinates (x, y) in meters within a global coordinate system and the vehicle's orientation in radians
relative to the global frame.
Together, this comprehensive sensor suite enables detailed monitoring and analysis of the dynamic behaviors of all traffic
agents.  Your task is to leverage this data to identify and interpret the temporal actions and maneuvers of each agent
within the traffic environment.
The following are driving maneuvers and actions along with their respective descriptions:
Ego stopping:  Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to
comply with traffic rules and comes to a complete stop.
Ego turning left:  Ego is executing a left turn at an intersection or junction.
Ego performing u-turn:  Ego is performing a 180-degree turn at an intersection or junction, reversing its direction of
travel.
Ego turning right:  Ego is executing a right turn at an intersection or junction.
Ego decelerating:  Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or
to comply with traffic rules, without coming to a complete stop.
Your input consists of sequential data, captured over 6 frames and 6 seconds with the described sensor suite.
Ego:
Frame number:  0
x:  0.00
y:  0.00
rotation:  0.00
.
.
.
Frame number:  5
x:  7.75
y:  3.50
rotation:  0.82
Which of the following options best describes ego driving maneuver?  Options:
A. Ego stopping
B. Ego turning left
C. Ego performing u-turn
D. Ego turning right
E. Ego decelerating
Please answer only with the letter of an option from the multiple choice list, e.g.  A or B or C or D, and nothing else.
```

```
Answer:
```

```
GPT-4o [46]:  B                    DeepSeek V3 [12]:  D                    Llama 3.2 [1]:  A                    Correct:  B
```

Figure 23: LLM Ego scenario prompt

System Prompt:

```
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers
performed by various agents in diverse driving scenarios.
```

User Prompt:

```
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers
performed by various agents in diverse driving scenarios.|An agent refers to any participant in the traffic environment,
including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles.
The traffic environment is captured using a sophisticated suite of sensors mounted on an ego vehicle, which is a standard
passenger car.
The sensor suite includes a Light Detection and Ranging (LiDAR) sensor labeled LIDAR_TOP, mounted on the roof of the car.
This LiDAR sensor provides high-precision 3D spatial data about the surrounding environment.  LiDAR data is crucial for
precise spatial positioning and size, which helps in differentiating vehicle types and detecting movement.  For each agent,
the LiDAR data includes the frame number, the center position of the agent (x, y, z) in meters relative to the LiDAR, and a
heading angle in degrees relative to the LiDAR, and size dimensions such as width, length, and height in meters.
In addition to the LiDAR, six cameras are strategically positioned around the vehicle to provide a 360-degree field of view.
Camera data can aid in visual confirmation and potentially understanding intent.  The first camera, CAM_FRONT, is positioned
directly in front of the LiDAR sensor and faces forward.  To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented
at a 45-degree angle relative to the front-facing camera.  On the right side of the car, CAM_BACK_RIGHT is positioned at a
135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the
left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera.  Finally,
CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. For each
agent, the camera data includes the frame number, the center of the agent given with pixel location (x, y) in the image
frame and in which camera is agent visible (e.g.  CAM_FRONT or CAM_BACK).
The system also integrates a GPS sensor, which provides the ego vehicle's precise global position.  The GPS data includes
the frame number, coordinates (x, y) in meters within a global coordinate system and the vehicle's orientation in radians
relative to the global frame.
Together, this comprehensive sensor suite enables detailed monitoring and analysis of the dynamic behaviors of all traffic
agents.  Your task is to leverage this data to identify and interpret the temporal actions and maneuvers of each agent
within the traffic environment.
The following are driving maneuvers and actions along with their respective descriptions:
Agent 1 leading ego:  Agent 1 travels ahead of ego at a similar speed while maintaining a consistent distance.
Ego overtaking agent 2:  Ego in the adjacent lane and moves ahead of agent 2 while both are in motion.
Ego decelerating:  Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or
to comply with traffic rules, without coming to a complete stop.
Ego waiting for pedestrian to cross:  Ego comes to a stop or remains stationary, yielding the right-of-way to agent 2 who is
crossing or preparing to cross the road, while maintaining awareness of the agent 2's movement and ensuring a safe distance
until the crossing is complete.
Ego passes stationary agent 2:  Ego in the adjacent lane overtakes the stopped agent 2.
Your input consists of sequential data, captured over 6 frames and 6 seconds with the described sensor suite.
Ego:
Frame number:  0
x:  0.00
y:  0.00
rotation:  0.00
.
.
.
Frame number:  5
x:  0.79
y:  0.03
rotation:  0.04
Agent 2:
Frame number:  0
LiDAR x:  -2.97
LiDAR y:  7.03
LiDAR rotation:  -2.48
CAM x:  221.21
CAM y:  630.17
CAM: CAM_FRONT
.
.
.
Frame number:  5
LiDAR x:  -5.41
LiDAR y:  4.08
LiDAR rotation:  -2.30
CAM x:  828.79
CAM y:  578.67
CAM: CAM_FRONT_LEFT
Which of the following options best describes ego driving behavior with respect to agent 2?
Options:
A. Agent 1 leading ego
B. Ego overtaking agent 2
C. Ego decelerating
D. Ego waiting for pedestrian to cross
E. Ego passes stationary agent 2
Please answer only with the letter of an option from the multiple choice list, e.g.  A or B or C or D, and nothing else.
```

Answer:

GPT-4o [46]: C	DeepSeek V3 [12]: C	Llama 3.2 [1]: B	Correct: D

Figure 24: LLM Ego-to-Agent scenario prompt

```
System Prompt:
```

```
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers
performed by various agents in diverse driving scenarios.
```

```
User Prompt:
```

```
An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction
vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles.  The traffic environment is captured using a
sophisticated suite of sensors mounted on an ego vehicle, which is a standard passenger car.
The sensor suite includes a Light Detection and Ranging (LiDAR) sensor labeled LIDAR_TOP, mounted on the roof of the car.
This LiDAR sensor provides high-precision 3D spatial data about the surrounding environment.  LiDAR data is crucial for
precise spatial positioning and size, which helps in differentiating vehicle types and detecting movement.  For each agent,
the LiDAR data includes the frame number, the center position of the agent (x, y, z) in meters relative to the LiDAR, and a
heading angle in degrees relative to the LiDAR, and size dimensions such as width, length, and height in meters.
In addition to the LiDAR, six cameras are strategically positioned around the vehicle to provide a 360-degree field of view.
Camera data can aid in visual confirmation and potentially understanding intent.  The first camera, CAM_FRONT, is positioned
directly in front of the LiDAR sensor and faces forward.  To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented
at a 45-degree angle relative to the front-facing camera.  On the right side of the car, CAM_BACK_RIGHT is positioned at a
135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the
left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera.  Finally,
CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. For each
agent, the camera data includes the frame number, the center of the agent given with pixel location (x, y) in the image
frame and in which camera is agent visible (e.g.  CAM_FRONT or CAM_BACK).
The system also integrates a GPS sensor, which provides the ego vehicle's precise global position.  The GPS data includes
the frame number, coordinates (x, y) in meters within a global coordinate system and the vehicle's orientation in radians
relative to the global frame.
Together, this comprehensive sensor suite enables detailed monitoring and analysis of the dynamic behaviors of all traffic
agents.  Your task is to leverage this data to identify and interpret the temporal actions and maneuvers of each agent
within the traffic environment.
The following are driving maneuvers and actions along with their respective descriptions:
Agent 1 turning left:  Agent 1 is executing a left turn at an intersection or junction.
Agent 1 changing lanes:  Agent 1 is transitioning from its current lane to an adjacent lane.
Agent 1 reversing:  Agent 1 is moving in reverse, either to park, navigate a tight space, or adjust its position.
Agent 1 decelerating:  Agent 1 is reducing its speed, either gradually or abruptly, in response to traffic conditions,
obstacles, or to comply with traffic rules, without coming to a complete stop.
Agent 1 performing u-turn:  Agent 1 is performing a 180-degree turn at an intersection or junction, reversing its direction
of travel.
Your input consists of sequential data, captured over 6 frames and 6 seconds with the described sensor suite.
Ego:
Frame number:  0
x:  0.00
y:  0.00
rotation:  0.00
.
.
Frame number:  5
x:  15.50
y:  -0.17
rotation:  -0.01
Agent 1:
Class:  car
Frame number:  0
LiDAR x:  5.17
LiDAR y:  -5.98
LiDAR rotation:  1.47
CAM x:  1326.79
CAM y:  643.39
CAM: CAM_BACK_RIGHT
.
.
Frame number:  5
LiDAR x:  5.91
LiDAR y:  -11.37
LiDAR rotation:  1.52
CAM x:  1599.80
CAM y:  480.51
CAM: CAM_BACK_RIGHT
Which of the following options best describes agent 1 driving maneuver?
Options:
A. Agent 1 turning left
B. Agent 1 changing lanes
C. Agent 1 reversing
D. Agent 1 decelerating
E. Agent 1 performing u-turn
Please answer only with the letter of an option from the multiple choice list, e.g.  A or B or C or D, and nothing else.
```

```
Answer:
```

```
GPT-4o [46]:  C.                DeepSeek V3 [12]:  B                Llama 3.2 [1]:  B                Correct:  B
```

Figure 25: LLM Agent scenario prompt

System Prompt:

```
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers
performed by various agents in diverse driving scenarios.
```

User Prompt:

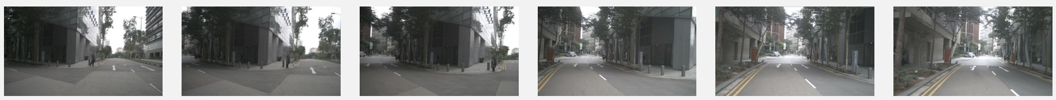
```
An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction
vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles.  The traffic environment is captured using a
sophisticated suite of sensors mounted on an ego vehicle, which is a standard passenger car.
The sensor suite includes a Light Detection and Ranging (LiDAR) sensor labeled LIDAR_TOP, mounted on the roof of the car.
This LiDAR sensor provides high-precision 3D spatial data about the surrounding environment.  LiDAR data is crucial for
precise spatial positioning and size, which helps in differentiating vehicle types and detecting movement.  For each agent,
the LiDAR data includes the frame number, the center position of the agent (x, y, z) in meters relative to the LiDAR, and a
heading angle in degrees relative to the LiDAR, and size dimensions such as width, length, and height in meters.
In addition to the LiDAR, six cameras are strategically positioned around the vehicle to provide a 360-degree field of view.
Camera data can aid in visual confirmation and potentially understanding intent.  The first camera, CAM_FRONT, is positioned
directly in front of the LiDAR sensor and faces forward.  To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented
at a 45-degree angle relative to the front-facing camera.  On the right side of the car, CAM_BACK_RIGHT is positioned at a
135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the
left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera.  Finally,
CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. For each
agent, the camera data includes the frame number, the center of the agent given with pixel location (x, y) in the image
frame and in which camera is agent visible (e.g.  CAM_FRONT or CAM_BACK).
The system also integrates a GPS sensor, which provides the ego vehicle's precise global position.  The GPS data includes
the frame number, coordinates (x, y) in meters within a global coordinate system and the vehicle's orientation in radians
relative to the global frame.
Together, this comprehensive sensor suite enables detailed monitoring and analysis of the dynamic behaviors of all traffic
agents.  Your task is to leverage this data to identify and interpret the temporal actions and maneuvers of each agent
within the traffic environment.
The following are driving maneuvers and actions along with their respective descriptions:
Agent 1 stationary to the right of agent 2:  Agent 1 is fully stopped and remains stationary to the right of agent 2, which
is also stationary, such as when waiting at a traffic light or in a parking lot.
Agent 1 overtaking agent 2:  Agent 1 in the adjacent lane and moves ahead of agent 2 while both are in motion.
Agent 1 stationary in front of agent 2:  Agent 1 is fully stopped and remains stationary in front of agent 2, such as when
waiting at a traffic light, in a parking lot, or any other situation requiring queuing.
Agent 1 stationary to the left of agent 2:  Agent 1 is fully stopped and remains stationary to the left of agent 2, which is
also stationary, such as when waiting at a traffic light or in a parking lot.
Agent 1 moving to the right of agent 2:  Agent 1 is traveling in parallel to the right of agent 2 (e.g., in adjacent
lanes or side by side), with one vehicle maintaining a rightward offset relative to the other.  This could occur during
lane-matched driving on a multi-lane road or synchronized movement from a traffic light.
Your input consists of sequential data, captured over 6 frames and 6 seconds with the described sensor suite.
Ego:
.
.
.
Agent 1:
Class:  car
Frame number:  0
LiDAR x:  -11.89
LiDAR y:  28.12
LiDAR rotation:  -1.58
CAM x:  291.60
CAM y:  546.14
CAM: CAM_FRONT
.
.
.
Frame number:  5
LiDAR x:  -12.35
LiDAR y:  -27.44
LiDAR rotation:  -1.57
CAM x:  1223.86
CAM y:  520.56
CAM: CAM_BACK
Agent 2:
Class:  car
Frame number:  0
LiDAR x:  -15.65
LiDAR y:  26.08
LiDAR rotation:  -1.58
CAM x:  90.13
CAM y:  564.20
CAM: CAM_FRONT
.
.
.
Frame number:  5
LiDAR x:  -15.99
LiDAR y:  -26.44
LiDAR rotation:  -1.52
CAM x:  1351.97
CAM y:  527.46
CAM: CAM_BACK
Which of the following options best describes agent 1 driving behaviour with respect to agent 2?
Options:
A. Agent 1 stationary to the right of agent 2
B. Agent 1 overtaking agent 2
C. Agent 1 stationary in front of agent 2
D. Agent 1 stationary to the left of agent 2
E. Agent 1 moving to the right of agent 2
Please answer only with the letter of an option from the multiple choice list, e.g.  A or B or C or D, and nothing else.
```

Answer:

GPT-4o [46]: B	DeepSeek V3 [12]: E	Llama 3.2 [1]: E	Correct: B

Figure 26: LLM Agent-to-Agent scenario prompt

Video:



Prompt:

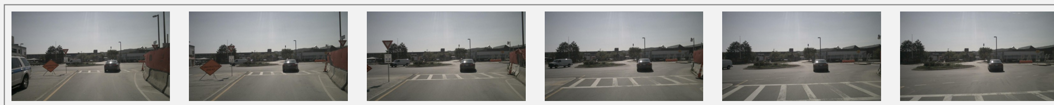
```
Frame-1:  IMAGE_TOKEN
Frame-2:  IMAGE_TOKEN
Frame-3:  IMAGE_TOKEN
Frame-4:  IMAGE_TOKEN
Frame-5:  IMAGE_TOKEN
Frame-6:  IMAGE_TOKEN
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers
performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios.  An agent
refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles,
trucks, trailers, motorcycles, pedestrians, and bicycles.  Your task is to identify both your temporal actions and maneuvers
and those of other agents within the traffic environment.
Six cameras are strategically positioned around the vehicle to provide a 360-degree field of view.  Camera data can aid in
visual confirmation and potentially understanding intent.  The first camera, CAM_FRONT, is positioned directly in front
of the LiDAR sensor and faces forward.  To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree
angle relative to the front-facing camera.  On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree
angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left
of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera.  Finally,
CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. You are
provided the six sequential video frames captured at 2 frames per second.
The following are driving maneuvers and actions along with their respective descriptions:
You are decelerating:  You are reducing your speed, either gradually or abruptly, in response to traffic conditions,
obstacles, or to comply with traffic rules, without coming to a complete stop.
You are changing lanes:  You are transitioning from your current lane to an adjacent lane.
You are turning right:  You are executing a right turn at an intersection or junction.
You are stopping:  You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles,
or to comply with traffic rules and comes to a complete stop.
You are turning left:  You are executing a left turn at an intersection or junction.
Given that, Frame-1 is captured with CAM_FRONT, Frame-2 is captured with CAM_FRONT, Frame-3 is caputred with CAM_FRONT,
Frame-4 is captured with CAM_FRONT, Frame-5 is captured with CAM_FRONT, Frame-6 is captured with CAM_FRONT, which of the
following options best describes your driving maneuver?
Options:
A. You are decelerating
B. You are changing lanes
C. You are turning right
D. You are stopping
E. You are turning left
Please answer only with the letter of an option from the multiple choice list, e.g.  A or B or C or D, and nothing else.
```

Answer:

```
InternVL 2.5 1B [9]:  A. You are decelerating        InternVL 2.5 8B [9]:  E        Correct:  E
```

Figure 27: VLM InternVL 2.5 8B/1B Ego scenario prompt

Video:



Prompt:

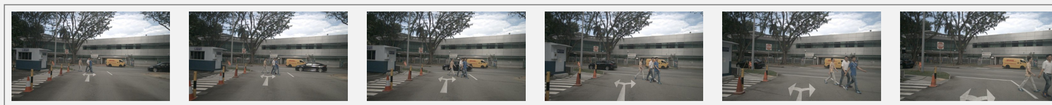
Frame-1: IMAGE_TOKEN
Frame-2: IMAGE_TOKEN
Frame-3: IMAGE_TOKEN
Frame-4: IMAGE_TOKEN
Frame-5: IMAGE_TOKEN
Frame-6: IMAGE_TOKEN
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.
Six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally, CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. You are provided the six sequential video frames captured at 2 frames per second.
The following are driving maneuvers and actions along with their respective descriptions:
You are stationary to the right of object 2: You are fully stopped and remain stationary to the right of object 2, which is also stationary, such as when waiting at a traffic light or in a parking lot.
You are overtaking object 2: You are the adjacent lane and move ahead of object 2 while both are in motion.
You are stopping: You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.
You are following object 2: You are driving behind object 2 at a similar speed while maintaining a consistent distance.
You are stationary behind object 2: Object 1 are fully stopped and remain stationary behind object 2, such as when waiting at a traffic light, in a parking lot, or any other situation requiring queuing.
Consider that the Frame-1 is captured with CAM_FRONT, Frame-2 is captured with CAM_FRONT, Frame-3 is caputred with CAM_FRONT, Frame-4 is captured with CAM_FRONT, Frame-5 is captured with CAM_FRONT, Frame-6 is captured with CAM_FRONT. Also, consider object 2, <ref>car</ref><box>[578, 523, 688, 682]</box> in Frame-1. Which of the following options best describes your driving behavior with respect to the object 2?
Options:
A. You are stationary to the right of object 2
B. You are overtaking object 2
C. You are stopping
D. You are following object 2
E. You are stationary behind object 2
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

InternVL 2.5 1B [9]: E. You are stationary behind object 2 InternVL 2.5 8B [9]: D Correct: D

Figure 28: VLM InternVL 2.5 8B/1B Ego-to-Agent scenario prompt

Video:



Prompt:

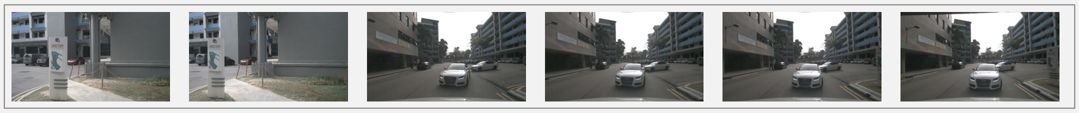
```
Frame-1: IMAGE_TOKEN
Frame-2: IMAGE_TOKEN
Frame-3: IMAGE_TOKEN
Frame-4: IMAGE_TOKEN
Frame-5: IMAGE_TOKEN
Frame-6: IMAGE_TOKEN
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers
performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent
refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles,
trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers
and those of other agents within the traffic environment.
Six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in
visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front
of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree
angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree
angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left
of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally,
CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. You are
provided the six sequential video frames captured at 2 frames per second.
The following are driving maneuvers and actions along with their respective descriptions:
Object 1 is decelerating: Object 1 is reducing its speed, either gradually or abruptly, in response to traffic conditions,
obstacles, or to comply with traffic rules, without coming to a complete stop.
Object 1 is accelerating: Object 1 is increasing its speed, either gradually or abruptly, to adapt to traffic conditions,
maintain flow, or comply with traffic rules and signals.
Object 1 is running: Object 1 (pedestrian) is running and moves rapidly.
Object 1 is crossing street: Object 1 (pedestrian) moves from one side of the road to the other, at a designated crossing
point or intersection.
Object 1 is jaywalking: Object 1 (pedestrian) crosses the street outside of designated crossing areas or against
traffic signals, often requiring heightened awareness of vehicle movements, quick decision-making to avoid conflicts, and
potentially creating unpredictable interactions with other agents in the traffic environment.
Consider that the Frame-1 is captured with CAM_FRONT, Frame-2 is captured with CAM_FRONT, Frame-3 is caputred with CAM_FRONT,
Frame-4 is captured with CAM_FRONT, Frame-5 is captured with CAM_FRONT, Frame-6 is captured with CAM_FRONT. Also, consider
object 1, <ref>pedestrian</ref><box>[443, 510, 519, 701]</box> in Frame-1. Which of the following options best describes
object 1 maneuver?
Options:
A. Object 1 is decelerating
B. Object 1 is accelerating
C. Object 1 is running
D. Object 1 is crossing street
E. Object 1 is jaywalking
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.
```

Answer:

```
InternVL 2.5 1B [9]: D          InternVL 2.5 8B [9]: D          Correct: E
```

Figure 29: VLM InternVL 2.5 8B/1B Agent scenario prompt

Video:



Prompt:

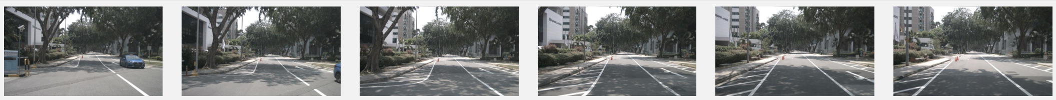
Frame-1: IMAGE_TOKEN
Frame-2: IMAGE_TOKEN
Frame-3: IMAGE_TOKEN
Frame-4: IMAGE_TOKEN
Frame-5: IMAGE_TOKEN
Frame-6: IMAGE_TOKEN
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.
Six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally, CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. You are provided the six sequential video frames captured at 2 frames per second.
The following are driving maneuvers and actions along with their respective descriptions:
Object 1 is following object 2: Object 1 is driving behind object 2 at a similar speed while maintaining a consistent distance.
Object 1 is stationary to the left of object 2: Object 1 is fully stopped and remains stationary to the left of object 2, which is also stationary, such as when waiting at a traffic light or in a parking lot.
Object 1 is accelerating: Object 1 is increasing its speed, either gradually or abruptly, to adapt to traffic conditions, maintain flow, or comply with traffic rules and signals.
Object 1 is moving to the right of object 2: Object 1 is traveling in parallel to the right of object 2 (e.g., in adjacent lanes or side by side), with one vehicle maintaining a rightward offset relative to the other. This could occur during lane-matched driving on a multi-lane road or synchronized movement from a traffic light.
Object 1 is leading object 2: Object 1 travels ahead of object 2 at a similar speed while maintaining a consistent distance.
Consider that the Frame-1 is captured with CAM_BACK_LEFT, Frame-2 is captured with CAM_BACK_LEFT, Frame-3 is captured with CAM_BACK, Frame-4 is captured with CAM_BACK, Frame-5 is captured with CAM_BACK, Frame-6 is captured with CAM_BACK. Also, consider object 1 <ref>car</ref><box>[71, 467, 251, 608]</box> in Frame-1 and object 2 <ref>car</ref><box>[150, 491, 321, 615]</box> in Frame-1. Which of the following options best describes object 1 maneuver with respect to the object 2?
Options:
A. Object 1 is following object 2
B. Object 1 is stationary to the left of object 2
C. Object 1 is accelerating
D. Object 1 is moving to the right of object 2
E. Object 1 is leading object 2
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

InternVL 2.5 1B [9]: C. Object 1 is accelerating InternVL 2.5 8B [9]: B Correct: B

Figure 30: VLM InternVL 2.5 8B/1B Agent-to-Agent scenario prompt

Video:



Prompt:

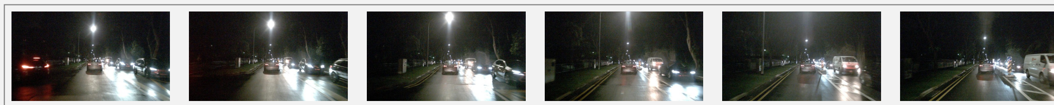
```
Frame-1: {IMAGE_TOKEN}
Frame-2: {IMAGE_TOKEN}
Frame-3: {IMAGE_TOKEN}
Frame-4: {IMAGE_TOKEN}
Frame-5: {IMAGE_TOKEN}
Frame-6: {IMAGE_TOKEN}
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers
performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent
refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles,
trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers
and those of other agents within the traffic environment.
Six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in
visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front
of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree
angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree
angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left
of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally,
CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. You are
provided the six sequential video frames captured at 2 frames per second.
The following are driving maneuvers and actions along with their respective descriptions:
You are stopping: You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles,
or to comply with traffic rules and comes to a complete stop.
You are accelerating: You are increasing your speed, either gradually or abruptly, to adapt to traffic conditions, maintain
flow, or comply with traffic rules and signals.
You are performing u-turn: You are performing a 180-degree turn at an intersection or junction, reversing its direction of
travel.
You are turning left: You are executing a left turn at an intersection or junction.
You are turning right: You are executing a right turn at an intersection or junction.
Given that, Frame-1 is captured with CAM_FRONT, Frame-2 is captured with CAM_FRONT, Frame-3 is caputred with CAM_FRONT,
Frame-4 is captured with CAM_FRONT, Frame-5 is captured with CAM_FRONT, Frame-6 is captured with CAM_FRONT, which of the
following options best describes your driving maneuver?
Options:
A. You are stopping
B. You are accelerating
C. You are performing u-turn
D. You are turning left
E. You are turning right
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.
```

Answer:

Qwen2.5-VL 7B [5]: D   Correct: B

Figure 31: VLM Qwen 2.5 7B Ego scenario prompt

Video:



Prompt:

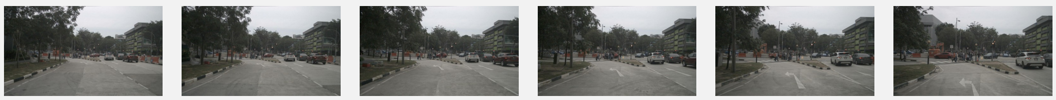
```
Frame-1: {IMAGE_TOKEN}
Frame-2: {IMAGE_TOKEN}
Frame-3: {IMAGE_TOKEN}
Frame-4: {IMAGE_TOKEN}
Frame-5: {IMAGE_TOKEN}
Frame-6: {IMAGE_TOKEN}
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers
performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent
refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles,
trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers
and those of other agents within the traffic environment.
Six cameras are strategically positioned around the vehicle to provide a 360-degree field of view. Camera data can aid in
visual confirmation and potentially understanding intent. The first camera, CAM_FRONT, is positioned directly in front
of the LiDAR sensor and faces forward. To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree
angle relative to the front-facing camera. On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree
angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left
of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera. Finally,
CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. You are
provided the six sequential video frames captured at 2 frames per second.
The following are driving maneuvers and actions along with their respective descriptions:
You are stationary behind object 2: Object 1 are fully stopped and remain stationary behind object 2, such as when waiting
at a traffic light, in a parking lot, or any other situation requiring queuing.
You are following object 2: You are driving behind object 2 at a similar speed while maintaining a consistent distance.
You are decelerating: You are reducing your speed, either gradually or abruptly, in response to traffic conditions,
obstacles, or to comply with traffic rules, without coming to a complete stop.
You are accelerating: You are increasing your speed, either gradually or abruptly, to adapt to traffic conditions, maintain
flow, or comply with traffic rules and signals.
You are passing stationary object 2: You in the adjacent lane overtakes the stopped object 2.
Consider that the Frame-1 is captured with CAM_FRONT, Frame-2 is captured with CAM_FRONT, Frame-3 is caputred with CAM_FRONT,
Frame-4 is captured with CAM_FRONT, Frame-5 is captured with CAM_FRONT, Frame-6 is captured with CAM_FRONT. Also, consider
object 2, which is car inside region [370, 233, 459, 307] in Frame-1. Which of the following options best describes your
driving behavior with respect to the object 2?
Options:
A. You are stationary behind object 2
B. You are following object 2
C. You are decelerating
D. You are accelerating
E. You are passing stationary object 2
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.
```

Answer:

```
Qwen2.5-VL 7B [5]: B                                                                                    Correct: B
```

Figure 32: VLM Qwen 2.5 7B Ego-to-Agent scenario prompt

Video:



Prompt:

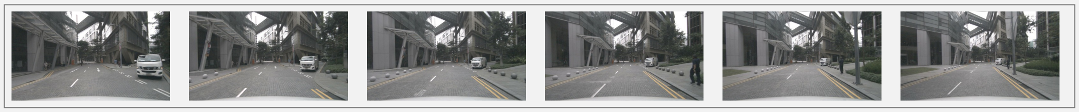
```
Frame-1:  {IMAGE_TOKEN}
Frame-2:  {IMAGE_TOKEN}
Frame-3:  {IMAGE_TOKEN}
Frame-4:  {IMAGE_TOKEN}
Frame-5:  {IMAGE_TOKEN}
Frame-6:  {IMAGE_TOKEN}
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers
performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios.  An agent
refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles,
trucks, trailers, motorcycles, pedestrians, and bicycles.  Your task is to identify both your temporal actions and maneuvers
and those of other agents within the traffic environment.
Six cameras are strategically positioned around the vehicle to provide a 360-degree field of view.  Camera data can aid in
visual confirmation and potentially understanding intent.  The first camera, CAM_FRONT, is positioned directly in front
of the LiDAR sensor and faces forward.  To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree
angle relative to the front-facing camera.  On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree
angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left
of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera.  Finally,
CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. You are
provided the six sequential video frames captured at 2 frames per second.
The following are driving maneuvers and actions along with their respective descriptions:
Object 1 is crossing street:  Object 1 (pedestrian) moves from one side of the road to the other, at a designated crossing
point or intersection.
Object 1 is accelerating:  Object 1 is increasing its speed, either gradually or abruptly, to adapt to traffic conditions,
maintain flow, or comply with traffic rules and signals.
Object 1 is decelerating:  Object 1 is reducing its speed, either gradually or abruptly, in response to traffic conditions,
obstacles, or to comply with traffic rules, without coming to a complete stop.
Object 1 is stationary:  Object 1 (pedestrian) remains stationary in the traffic environment, either waiting at a crossing,
observing surroundings, or pausing for other reasons.
Object 1 is jaywalking:  Object 1 (pedestrian) crosses the street outside of designated crossing areas or against
traffic signals, often requiring heightened awareness of vehicle movements, quick decision-making to avoid conflicts, and
potentially creating unpredictable interactions with other agents in the traffic environment.
Consider that the Frame-1 is captured with CAM_FRONT, Frame-2 is captured with CAM_FRONT, Frame-3 is caputred with CAM_FRONT,
Frame-4 is captured with CAM_FRONT, Frame-5 is captured with CAM_FRONT, Frame-6 is captured with CAM_FRONT. Also, consider
object 1, which is pedestrian inside region [296, 232, 313, 265] in Frame-1.  Which of the following options best describes
object 1 maneuver?
Options:
A. Object 1 is crossing street
B. Object 1 is accelerating
C. Object 1 is decelerating
D. Object 1 is stationary
E. Object 1 is jaywalking
Please answer only with the letter of an option from the multiple choice list, e.g.  A or B or C or D, and nothing else.
```

Answer:

Qwen2.5-VL 7B [5]: A   Correct: A

Figure 33: VLM Qwen 2.5 7B Agent scenario prompt

Video:



Prompt:

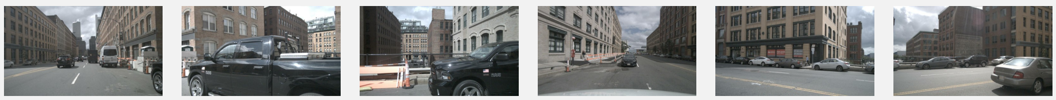
```
Frame-1:  {IMAGE_TOKEN}
Frame-2:  {IMAGE_TOKEN}
Frame-3:  {IMAGE_TOKEN}
Frame-4:  {IMAGE_TOKEN}
Frame-5:  {IMAGE_TOKEN}
Frame-6:  {IMAGE_TOKEN}
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers
performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios.  An agent
refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles,
trucks, trailers, motorcycles, pedestrians, and bicycles.  Your task is to identify both your temporal actions and maneuvers
and those of other agents within the traffic environment.
Six cameras are strategically positioned around the vehicle to provide a 360-degree field of view.  Camera data can aid in
visual confirmation and potentially understanding intent.  The first camera, CAM_FRONT, is positioned directly in front
of the LiDAR sensor and faces forward.  To the right of CAM_FRONT is CAM_FRONT_RIGHT, which is oriented at a 45-degree
angle relative to the front-facing camera.  On the right side of the car, CAM_BACK_RIGHT is positioned at a 135-degree
angle relative to CAM_FRONT. The rear-facing camera, CAM_BACK, is oriented directly opposite to CAM_FRONT. To the left
of CAM_FRONT is CAM_FRONT_LEFT, which is oriented at a -45-degree angle relative to the front-facing camera.  Finally,
CAM_BACK_LEFT is positioned on the left side of the car, oriented at a -135-degree angle relative to CAM_FRONT. You are
provided the six sequential video frames captured at 2 frames per second.
The following are driving maneuvers and actions along with their respective descriptions:
Object 1 is running:  Object 1 (pedestrian) is running and moves rapidly.
Object 1 is overtaking object 2:  Object 1 in the adjacent lane and moves ahead of object 2 while both are in motion.
Object 1 is passing stationary object 2:  Object 1 in the adjacent lane overtakes the stopped object 2.
Object 1 is crossing street:  Object 1 (pedestrian) moves from one side of the road to the other, at a designated crossing
point or intersection.
Object 1 is walking alongside object 2:  Object 1 (pedestrian) and object 2 (pedestrian) walk side by side at a steady,
moderate pace.
Consider that the Frame-1 is captured with CAM_BACK, Frame-2 is captured with CAM_BACK, Frame-3 is caputred with CAM_BACK,
Frame-4 is captured with CAM_BACK, Frame-5 is captured with CAM_BACK, Frame-6 is captured with CAM_BACK. Also, consider
object 1, which is pedestrian inside region [510, 239, 536, 283] in Frame-1 and object 2, which is pedestrian inside region
[534, 238, 556, 280] in Frame-1.  Which of the following options best describes object 1 maneuver with respect to the object
2?
Options:
A. Object 1 is running
B. Object 1 is overtaking object 2
C. Object 1 is passing stationary object 2
D. Object 1 is crossing street
E. Object 1 is walking alongside object 2
Please answer only with the letter of an option from the multiple choice list, e.g.  A or B or C or D, and nothing else.
```

Answer:

Qwen2.5-VL 7B [5]: E   Correct: E

Figure 34: VLM Qwen 2.5 7B Agent-to-Agent scenario prompt

Multi-view Image Sequence:



Prompt:

A chat between a curious human and an artificial intelligence assistant. The assistent is specilized in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as you, across diverse driving scenarios. Agents refer to all participants in the traffic environment, including but not limited to: cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. You are the primary vehicle from whose perspective the scenario is being evaluated. You are equipped with a sophisticated suite of sensors (e.g., cameras, LiDAR, radar) to capture the surrounding traffic environment. Temporal actions and maneuvers include any time-based behaviors or movements, such as lane changes, accelerations, decelerations, turns, stops, or interactions between agents and between you and agents.
The following are driving maneuvers and actions along with their respective descriptions:
Decelerating: You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.
Stopping: You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.
Reversing: You are moving in reverse, either to park, navigate a tight space, or adjust your position.
Turning left: You are executing a left turn at an intersection or junction.
Turning right: You are executing a right turn at an intersection or junction.
The assistant gives helpful, detailed, and polite answers to the human's questions. USER: <FRONT VIEW>:
<image>
<FRONT LEFT VIEW>:
<image>
<FRONT RIGHT VIEW>:
<image>
<BACK LEFT VIEW>:
<image>
<BACK RIGHT VIEW>:
<image>
<BACK VIEW>:
<image>
You are driving, which of the following options best describes your driving maneuver?
Options:
A. Decelerating
B. Stopping
C. Reversing
D. Turning left
E. Turning right
For example, a correct answer format is like 'A'. ASSISTANT:

Answer:

Senna-VLM [29]: A. Decelerating Correct: E

Figure 35: Senna-VLM [29] Ego scenario prompt

Multi-view Image Sequence:

Prompt:

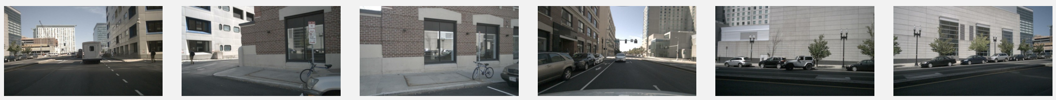
A chat between a curious human and an artificial intelligence assistant. The assistent is specilized in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as you, across diverse driving scenarios. Agents refer to all participants in the traffic environment, including but not limited to: cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. You are the primary vehicle from whose perspective the scenario is being evaluated. You are equipped with a sophisticated suite of sensors (e.g., cameras, LiDAR, radar) to capture the surrounding traffic environment. Temporal actions and maneuvers include any time-based behaviors or movements, such as lane changes, accelerations, decelerations, turns, stops, or interactions between agents and between you and agents.
The following are driving maneuvers and actions along with their respective descriptions:
You overtaking object 2: You are the adjacent lane and move ahead of object 2 while both are in motion.
Decelerating: You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.
You passes stationary object 2: You in the adjacent lane overtakes the stopped object 2.
You stationary in front of object 2: You are fully stopped and remain stationary in front of object 2, such as when waiting at a traffic light, in a parking lot, or any other situation requiring queuing.
Reversing: You are moving in reverse, either to park, navigate a tight space, or adjust your position.
The assistant gives helpful, detailed, and polite answers to the human's questions. USER: <FRONT VIEW>:
<image>
<FRONT LEFT VIEW>:
<image>
<FRONT RIGHT VIEW>:
<image>
<BACK LEFT VIEW>:
<image>
<BACK RIGHT VIEW>:
<image>
<BACK VIEW>:
<image>
I will now provide you with the position and velocity information of the dynamic objects:
Object 2: car, 7 meters ahead, 6 meters right, speed of 0 m/s.
Please predict which of the following options best describes your driving behavior with respect to Object 2.
Options:
A. You overtaking object 2
B. Decelerating
C. You passes stationary object 2
D. You stationary in front of object 2
E. Reversing
For example, a correct answer format is like 'A'. ASSISTANT:

Answer:

Senna-VLM [29]: A. You overtaking object 2 Correct: C

Figure 36: Senna-VLM [29] Ego-to-Agent scenario prompt

Multi-view Image Sequence:



Prompt:

A chat between a curious human and an artificial intelligence assistant. The assistent is specilized in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as you, across diverse driving scenarios. Agents refer to all participants in the traffic environment, including but not limited to: cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. You are the primary vehicle from whose perspective the scenario is being evaluated. You are equipped with a sophisticated suite of sensors (e.g., cameras, LiDAR, radar) to capture the surrounding traffic environment. Temporal actions and maneuvers include any time-based behaviors or movements, such as lane changes, accelerations, decelerations, turns, stops, or interactions between agents and between you and agents.
The following are driving maneuvers and actions along with their respective descriptions:
U-turn: Object 1 is performing a 180-degree turn at an intersection or junction, reversing its direction of travel.
Stopping: Object 1 is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.
Turning right: Object 1 is executing a right turn at an intersection or junction.
Reversing: Object 1 is moving in reverse, either to park, navigate a tight space, or adjust its position.
Accelerating: Object 1 is increasing its speed, either gradually or abruptly, to adapt to traffic conditions, maintain flow, or comply with traffic rules and signals.
The assistant gives helpful, detailed, and polite answers to the human's questions. USER: <FRONT VIEW>:
<image>
<FRONT LEFT VIEW>:
<image>
<FRONT RIGHT VIEW>:
<image>
<BACK LEFT VIEW>:
<image>
<BACK RIGHT VIEW>:
<image>
<BACK VIEW>:
<image>
I will now provide you with the position and velocity information of the dynamic objects:
Object 1: car, 0 meters behind, 14 meters left, speed of 1 m/s.
Please predict which of the following options best describes Object 1 driving behavior.
Options:
A. U-turn
B. Stopping
C. Turning right
D. Reversing
E. Accelerating
For example, a correct answer format is like 'A'. ASSISTANT:

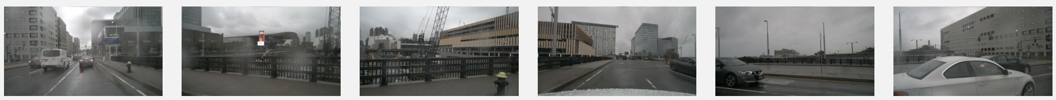
Answer:

Senna-VLM [29]: Object 1 is moving in the same direction as the camera, and there is no indication of a U-turn or a stop. The car is maintaining a safe distance from the vehicle in front, which is also moving in the same direction. The road appears to be clear, and there are no visible obstacles or traffic signals requiring immediate attention. The weather conditions seem to be clear and sunny, which is favorable for driving. The time of day appears to be daytime, as there is ample natural light. The overall driving environment is safe and conducive to maintaining a steady speed.

Correct: D

Figure 37: Senna-VLM [29] Agent scenario prompt

Multi-view Image Sequence:



Prompt:

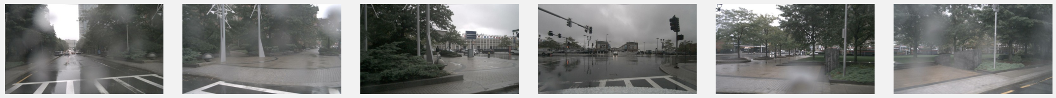
A chat between a curious human and an artificial intelligence assistant. The assistent is specilized in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as you, across diverse driving scenarios. Agents refer to all participants in the traffic environment, including but not limited to: cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. You are the primary vehicle from whose perspective the scenario is being evaluated. You are equipped with a sophisticated suite of sensors (e.g., cameras, LiDAR, radar) to capture the surrounding traffic environment. Temporal actions and maneuvers include any time-based behaviors or movements, such as lane changes, accelerations, decelerations, turns, stops, or interactions between agents and between you and agents.
The following are driving maneuvers and actions along with their respective descriptions:
Object 1 overtaking object 2: Object 1 in the adjacent lane and moves ahead of object 2 while both are in motion.
Turning right: Object 1 is executing a right turn at an intersection or junction.
Object 1 waiting for pedestrian to cross: Object 1 comes to a stop or remains stationary, yielding the right-of-way to a object 2 who is crossing or preparing to cross the road, while maintaining awareness of the object 2's movement and ensuring a safe distance until the crossing is complete.
Decelerating: Object 1 is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.
Object 1 stationary to the right of object 2: Object 1 is fully stopped and remains stationary to the right of object 2, which is also stationary, such as when waiting at a traffic light or in a parking lot.
The assistant gives helpful, detailed, and polite answers to the human's questions. USER: <FRONT VIEW>:
<image>
<FRONT LEFT VIEW>:
<image>
<FRONT RIGHT VIEW>:
<image>
<BACK LEFT VIEW>:
<image>
<BACK RIGHT VIEW>:
<image>
<BACK VIEW>:
<image>
I will now provide you with the position and velocity information of the dynamic objects:
Object 1: car, 15 meters ahead, 3 meters left, speed of 6 m/s.
Object 2: car, 18 meters ahead, 6 meters left, speed of 5 m/s.
Please predict which of the following options best describes Object 1 driving behavior with respect to Object 2.
Options:
A. Object 1 overtaking object 2
B. Turning right
C. Object 1 waiting for pedestrian to cross
D. Decelerating
E. Object 1 stationary to the right of object 2
For example, a correct answer format is like 'A'. ASSISTANT:

Answer:

Senna-VLM [29]: Object 1 is overtaking Object 2. Correct: A

Figure 38: Senna-VLM [29] Agent-to-Agent scenario prompt

Multi-view Image Sequence:



Prompt:

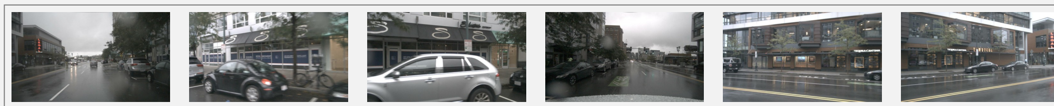
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.
The following are driving maneuvers and actions along with their respective descriptions:
You are changing lanes: You are transitioning from your current lane to an adjacent lane.
You are turning left: You are executing a left turn at an intersection or junction.
You are performing u-turn: You are performing a 180-degree turn at an intersection or junction, reversing its direction of travel.
You are decelerating: You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.
You are turning right: You are executing a right turn at an intersection or junction.
Which of the following options best describes your driving maneuver?
Options:
A. You are changing lanes
B. You are turning left
C. You are performing u-turn
D. You are decelerating
E. You are turning right
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

OmniDrive [60]: C Correct: A

Figure 39: OmniDrive [60] Ego scenario prompt

Multi-view Image Sequence:



Prompt:

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.
The following are driving maneuvers and actions along with their respective descriptions:
You are moving to the left of object 2: You are traveling in parallel to the left of object 2 (e.g., in adjacent lanes or side by side), with one vehicle maintaining a leftward offset relative to the other. This could occur during lane-matched driving on a multi-lane road or synchronized movement from a traffic light.
You are accelerating: You are increasing your speed, either gradually or abruptly, to adapt to traffic conditions, maintain flow, or comply with traffic rules and signals.
You are overtaking object 2: You are the adjacent lane and move ahead of object 2 while both are in motion.
You are passing stationary object 2: You in the adjacent lane overtakes the stopped object 2.
You are decelerating: You are reducing your speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.
Consider the object 2, which is a car located at coordinates (+4.4, +7.5) and moving at a velocity of 0.1 m/s. Which of the following options best describes your driving behavior with respect to the object 2?
Options:
A. You are moving to the left of object 2
B. You are accelerating
C. You are overtaking object 2
D. You are passing stationary object 2
E. You are decelerating
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

OmniDrive [60]: A Correct: D

Figure 40: OmniDrive [60] Ego-to-Agent scenario prompt

Multi-view Image Sequence:

Prompt:

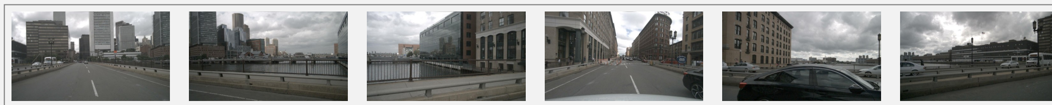
You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.
The following are driving maneuvers and actions along with their respective descriptions:
Object 1 is stopping: Object 1 is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.
Object 1 are changing lanes: Object 1 is transitioning from its current lane to an adjacent lane.
Object 1 is performing u-turn: Object 1 is performing a 180-degree turn at an intersection or junction, reversing its direction of travel.
Object 1 is turning right: Object 1 is executing a right turn at an intersection or junction.
Object 1 is turning left: Object 1 is executing a left turn at an intersection or junction.
Consider the object 1, which is a car located at coordinates (+0.3, -29.0) and moving at a velocity of 6.8 m/s. Which of the following options best describes object 1 maneuver?
Options:
A. Object 1 is stopping
B. Object 1 are changing lanes
C. Object 1 is performing u-turn
D. Object 1 is turning right
E. Object 1 is turning left
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

OmniDrive [60]: A | Correct: B

Figure 41: OmniDrive [60] Agent scenario prompt

Multi-view Image Sequence:



Prompt:

You are a helpful traffic control expert specializing in the analysis and identification of temporal actions and maneuvers performed by various agents, as well as your own temporal actions and maneuvers, in diverse driving scenarios. An agent refers to any participant in the traffic environment, including but not limited to cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify both your temporal actions and maneuvers and those of other agents within the traffic environment.
The following are driving maneuvers and actions along with their respective descriptions:
Object 1 is walking in opposite direction of object 2: Object 1 (pedestrian) and object 2 (pedestrian) walk toward each other at a moderate pace, cross paths, and proceed.
Object 1 is stationary: Object 1 (pedestrian) remains stationary in the traffic environment, either waiting at a crossing, observing surroundings, or pausing for other reasons.
Object 1 is jaywalking: Object 1 (pedestrian) crosses the street outside of designated crossing areas or against traffic signals, often requiring heightened awareness of vehicle movements, quick decision-making to avoid conflicts, and potentially creating unpredictable interactions with other agents in the traffic environment.
Object 1 is crossing street: Object 1 (pedestrian) moves from one side of the road to the other, at a designated crossing point or intersection.
Object 1 is passing stationary object 2: Object 1 in the adjacent lane overtakes the stopped object 2.
Consider the object 1, which is a pedestrian located at coordinates (-18.7, -2.5) and moving at a velocity of 1.8 m/s, and the object 2, whic is a pedestrian located at coordinates (-20.4, +1.1) and moving at a velocity of 1.3 m/s. Which of the following options best describes object 1 maneuver with respect to the object 2?
Options:
A. Object 1 is walking in opposite direction of object 2
B. Object 1 is stationary
C. Object 1 is jaywalking
D. Object 1 is crossing street
E. Object 1 is passing stationary object 2
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

OmniDrive [60]: E | Correct: A

Figure 42: OmniDrive [60] Agent-to-Agent scenario prompt

Multi-view Video:

Prompt:

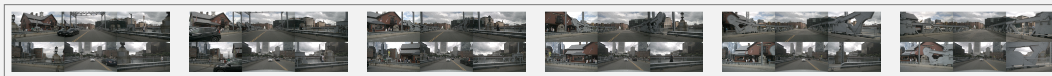
1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six images are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. You are a helpful traffic control expert specializing in analyzing and identifying the temporal actions and maneuvers of the ego vehicle and other agents in diverse driving scenarios. Agents include all traffic participants such as cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify the temporal actions and maneuvers of both ego and other agents within the traffic environment.
The following are driving maneuvers and actions along with their respective descriptions:
Ego is stopping: Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.
Ego is turning right: Ego is executing a right turn at an intersection or junction.
Ego is reversing: Ego is moving in reverse, either to park, navigate a tight space, or adjust its position.
Ego is decelerating: Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.
Ego is turning left: Ego is executing a left turn at an intersection or junction.
Which of the following options best describes ego driving maneuver?
Options:
A. Ego is stopping
B. Ego is turning right
C. Ego is reversing
D. Ego is decelerating
E. Ego is turning left
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

DriveMM [22]: B Correct: B

Figure 43: DriveMM [22] Ego scenario prompt

Multi-view Video:



Prompt:

1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six images are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. You are a helpful traffic control expert specializing in analyzing and identifying the temporal actions and maneuvers of the ego vehicle and other agents in diverse driving scenarios. Agents include all traffic participants such as cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify the temporal actions and maneuvers of both ego and other agents within the traffic environment.
The following are driving maneuvers and actions along with their respective descriptions:
Ego is decelerating: Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules, without coming to a complete stop.
Ego is following c2: Ego is driving behind c2 at a similar speed while maintaining a consistent distance.
Ego is stopping: Ego is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.
Ego is changing lanes: Ego is transitioning from its current lane to an adjacent lane.
Ego is leading c2: Ego travels ahead of c2 at a similar speed while maintaining a consistent distance.
Which of the following options best describes the ego driving behavior with respect to the <c2,CAM_BACK,54,59>?
Options:
A. Ego is decelerating
B. Ego is following c2
C. Ego is stopping
D. Ego is changing lanes
E. Ego is leading c2
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.

Answer:

DriveMM [22]: E Correct: B

Figure 44: DriveMM [22] Ego-to-Agent scenario prompt

Multi-view Video:

Prompt:

```
1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six images are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. You are a helpful traffic control expert specializing in analyzing and identifying the temporal actions and maneuvers of the ego vehicle and other agents in diverse driving scenarios. Agents include all traffic participants such as cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify the temporal actions and maneuvers of both ego and other agents within the traffic environment.
The following are driving maneuvers and actions along with their respective descriptions:
c1 is standing: c1 (pedestrian) remains stationary in the traffic environment, either waiting at a crossing, observing surroundings, or pausing for other reasons.
c1 is jaywalking: c1 (pedestrian) crosses the street outside of designated crossing areas or against traffic signals, often requiring heightened awareness of vehicle movements, quick decision-making to avoid conflicts, and potentially creating unpredictable interactions with other agents in the traffic environment.
c1 is crossing street: c1 (pedestrian) moves from one side of the road to the other, at a designated crossing point or intersection.
c1 is running: c1 (pedestrian) is running and moves rapidly.
c1 is walking: c1 (pedestrian) moves at a steady, moderate pace, typically following designated paths or crosswalks.
Which of the following options best describes the driving behavior of the <c1,CAM_FRONT,45,56>?
Options:
A. c1 is standing
B. c1 is jaywalking
C. c1 is crossing street
D. c1 is running
E. c1 is walking
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.
```

Answer:

DriveMM [22]: C   Correct: A

Figure 45: DriveMM [22] Agent scenario prompt

Multi-view Video:

Prompt:

```
1: <image> 2: <image> 3: <image> 4: <image> 5: <image> 6: <image>. These six images are the front view, front left view, front right view, back view, back left view and back right view of the ego vehicle. You are a helpful traffic control expert specializing in analyzing and identifying the temporal actions and maneuvers of the ego vehicle and other agents in diverse driving scenarios. Agents include all traffic participants such as cars, buses, construction vehicles, trucks, trailers, motorcycles, pedestrians, and bicycles. Your task is to identify the temporal actions and maneuvers of both ego and other agents within the traffic environment.
The following are driving maneuvers and actions along with their respective descriptions:
c1 is passes stationary c2: c1 in the adjacent lane overtakes the stopped c2.
c1 is stopping: c1 is reducing its speed, either gradually or abruptly, in response to traffic conditions, obstacles, or to comply with traffic rules and comes to a complete stop.
c1 is leading c2: c1 travels ahead of c2 at a similar speed while maintaining a consistent distance.
c1 is stationary to the right of c2: c1 is fully stopped and remains stationary to the right of c2, which is also stationary, such as when waiting at a traffic light or in a parking lot.
c1 is accelerating: c1 is increasing its speed, either gradually or abruptly, to adapt to traffic conditions, maintain flow, or comply with traffic rules and signals.
Which of the following options best describes <c1,CAM_FRONT_RIGHT,95,60> maneuver with respect to the <c2,CAM_FRONT_RIGHT,95,60>?
Options:
A. c1 is passes stationary c2
B. c1 is stopping
C. c1 is leading c2
D. c1 is stationary to the right of c2
E. c1 is accelerating
Please answer only with the letter of an option from the multiple choice list, e.g. A or B or C or D, and nothing else.
```

Answer:

DriveMM [22]: D   Correct: C

Figure 46: DriveMM [22] Agent-to-Agent scenario prompt