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

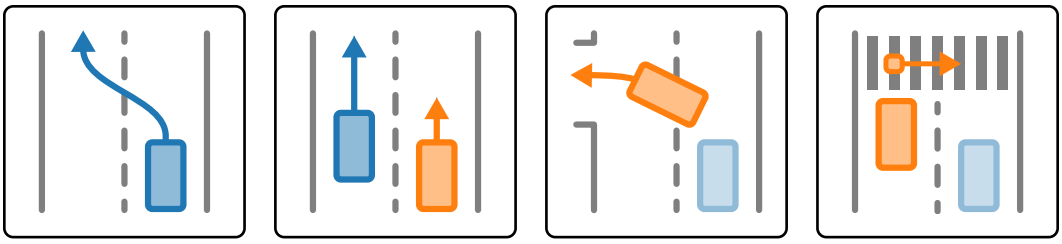
We introduce STSBench, a scenario-based framework to benchmark the holistic understanding of vision-language models (VLMs) for autonomous driving. The framework automatically mines predefined traffic scenarios from any dataset using ground-truth annotations, provides an intuitive user interface for efficient 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, focusing 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 advancements 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, 7, 8, 10, 31, 32, 37–39] has sparked significant interest in applying them to end-to-end autonomous driving [17, 23, 28, 33, 46, 49, 51, 57, 65, 67, 68]. These models aim to enhance trust in fully autonomous systems by providing human-interpretable decisions in natural language [60]. Unlike pretrained generalist

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

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.



(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 [28, 57, 68] 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 [26] simulated in CARLA [16]. A drawback of the open-loop evaluation on nuScenes [6], however, is the relatively small and unbalanced validation split [36] in which, for approximately 75% of all cases, the correct action is "continue to drive straight". Moreover, evaluating the 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 [19, 34, 41, 43, 48, 50, 53, 60, 64]. They typically focus on specific abilities such as spatial reasoning in camera images [19, 53], recognition and prediction of ego-vehicle actions [43, 50], the handling of visual corruptions [60], or understanding risk [41, 48] and critical driving scenarios [64]. 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 (agents). 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 [29, 30, 43, 48, 50, 54, 60] are typically tied to a specific dataset through manual annotations or extensive human verification. However, although many driving datasets [6, 42, 52, 58] 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 [35, 56] 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 agent-centric traffic scenarios, following a predefined scenario catalog, that reflect real-world interactions by leveraging information such as 3D bounding boxes and tracks, agent class labels, ego-motion data, and HD maps. The catalog contains all scenarios and their

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 [19]	S.I.	Auto	✓	✓	✗	✗	✗	MC, NUM	✓
NuScenes-MQA [24]	M.I.	Auto	✗	✓	✗	✗	✗	VQA	✓
NuScenes-QA [47]	M.I.	Auto	✗	✓	✗	✗	✗	VQA	✓
DriveLM [50]	M.I.	Manual	✓	✓	✗	✓	✓	VQA	✓
NuScenes-SpatialQA [53]	M.I.	Auto	✗	✓	✗	✗	✗	MC, NUM	✗
DriveBench [60]	M.I.	Manual	✗	✓	✗	✓	✓	VQA	✓
NuInstruct [15]	M.V.	Auto	✗	✓	✓	✓	✗	VQA	✓
TOD3Cap [29]	M.V.	Manual	✓	✓	✗	✗	✗	VQA	✓
DriveLLM-o1 [25]	M.I.	Auto	✓	✓	✓	✗	✗	VQA, MC	✓
OmniDrive [57]	M.V.	Auto	✓	✓	✓	✗	✗	VQA, OLP	✓
STSBench3D (ours)	M.V.	Auto	✓	✓	✓	✓	✓	MC	✓

definitions. Each scenario is also assigned a list of negative scenarios that most of the time should not occur concurrently (e.g., *left turn* and *overtaking*), representing negative answer candidates for the subsequent question generation. 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 (*i.e.*, mined but do not apply) and remove predefined negative ones that also occur during the corresponding scenario (*e.g.*, for a vehicle increasing its speed while performing a *right turn*, the negative scenario *accelerate* would be removed when *right turn* is the mined positive scenario). From the verified scenarios, STSBench automatically constructs a multiple-choice benchmark that asks models to identify which interactions occur in a given scene. In particular, the benchmark tests for the correct identification and understanding of agent-related spatio-temporal scenarios. Questions may concern the behavior of the ego vehicle, the actions of other agents (such as vehicles, pedestrians, or cyclists), interactions between the ego vehicle and other agents, or among 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 in text form, off-the-shelf VLMs operate on multiple images. Expert models are designed to deal with consecutive multi-view images (videos). 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, 9, 18, 42, 52, 58, 63] 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 [10, 37, 32], annotations have been added mostly for separate multi-view

images, focusing on tasks such as grounding [13, 55], ego-action prediction [30, 61], open-loop planning [50, 54], risk assessment [14, 41], spatial reasoning [19, 53], or visual question answering [43, 44, 34]. To address the lack of 3D understanding, EML [69] 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 [19], NuScenes-MQA [24], and NuScenes-SpatialQA [53]. 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 [15, 57] or 3D [47, 29, 59, 48] annotations emerged. Datasets such as Nuscenes-QA [47], DriveLM [50], OmniDrive [57], and NuInstruct [15] 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 [47, 50] and spatial relations [15], 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, 8, 12, 2, 32, 38, 10] 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, 40, 50, 62, 66, 69] 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 [28, 22, 33, 57]. To cope with the increasing number of image tokens caused by additional views and multiple video frames, Senna [28] 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 [33] and BEV-InMM [15] utilize a BEVFormer [35] backbone, OmniDrive [57] uses StreamPETR [56]. More recent approaches combine end-to-end driving models [20, 27] with VLMs [17, 46, 51, 68] 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 of the spatio-temporal reasoning capabilities of language-based end-to-end driving models by applying STSBench to instantiate STSnu on nuScenes [6].

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 [23, 28, 57] or extended existing planning models such as UniAD [20] with LLMs [68] 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 [19, 24, 53] or temporal [15, 25] reasoning, most are restricted to single frames or monocular views. NuInstruct [15] 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.

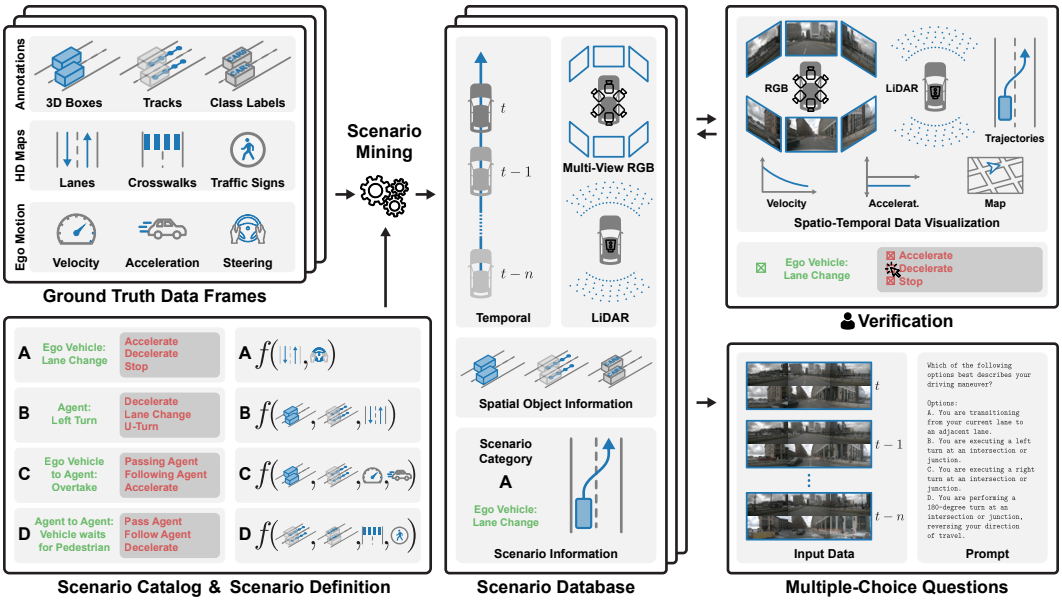


Figure 2: **STSBench workflow.** (Best viewed on screen)

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 [35, 56], commonly used in end-to-end driving models, learn this mapping 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 [29, 50], 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 ground truth perception and map annotations of any dataset 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 dataset-agnostic annotation framework aims to create an accurate benchmark for any given perception dataset (*e.g.*, [6, 42, 52]), considering all available ground truth annotations for the respective dataset. These annotations include 3D bounding boxes, tracking identifiers, agent class labels, ego-motion data (*e.g.*, velocity, acceleration, steering angle), or HD map data (*e.g.*, lanes, lane boundaries, 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.*). Additionally, we predefine a list of negative scenarios for each entry in the catalog. Negative scenarios are other scenarios that should not occur for the same agent concurrently and serve as wrong answer candidates in our question generation. Assume we have an *overtaking* scenario in which an agent (vehicle) in motion passes another agent (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

(assuming they have no semantic or temporal overlap) to test the reasoning capabilities: Holistic spatio-temporal understanding implies that also closely related scenarios can be distinguished.

2) Scenario definition. To enable automated mining of predefined traffic scenarios, we specify heuristics using only ground-truth annotations from the target dataset. 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 predefined scenarios and save them in a scenario database. The database contains references to visual data, such as consecutive images of all available views and 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 the quality of annotations. 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 agent scenario and verify that all assigned negative scenarios do not apply to the same agent. For instance, in a mined *lane change* scenario, the same agent might also reduce its speed (*deceleration*), which would consequently represent an invalid negative scenario if predefined in the scenario catalog (see Fig. 2, verification). In this example, the annotator would remove *deceleration* from the list of negative scenarios. 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 relevant agents. The questions are designed to have one correct answer and, by default, provide five possible choices. Further details are provided in the supplementary material.

3.2 STSnu Benchmark Construction

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

Data Source. Since most expert driving models [21, 29, 54, 57] 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 agent 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*. For control decisions and collision prevention, driving models must be aware of the ego vehicle's state 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 also includes 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 assessing risk, planning the next course

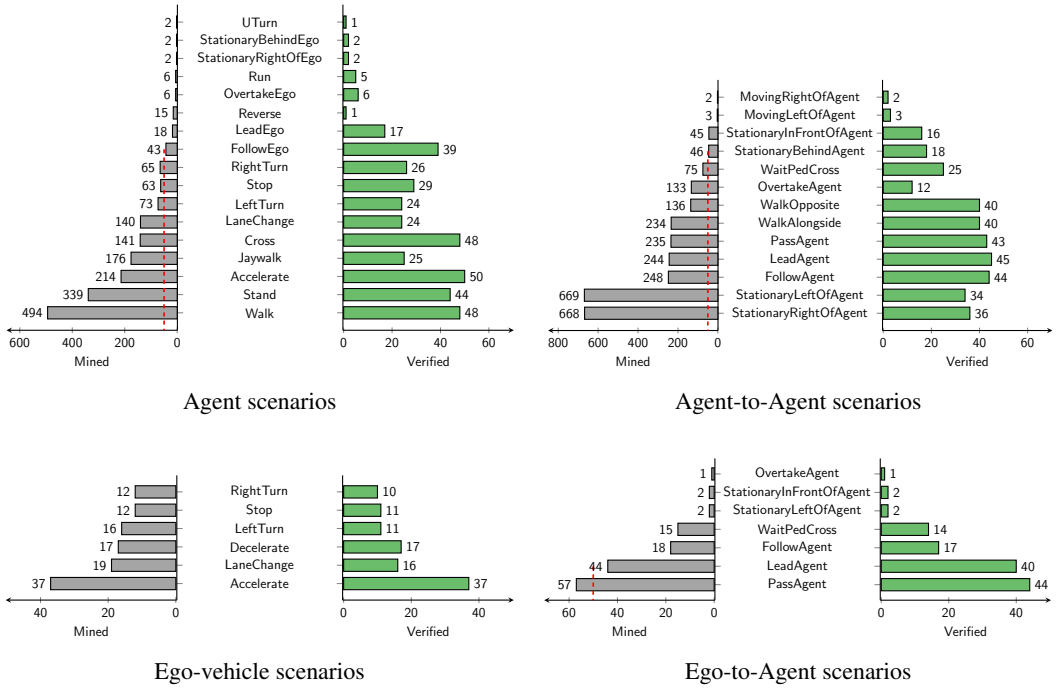


Figure 3: **Scenario statistics.** Number of mined scenarios in total (gray) and the remaining samples (green) after sub-sampling and verification. Scenarios with more than 50 samples (dashed red line) have been sub-sampled considering spatial distribution, occlusion, and distance to the ego-vehicle.

of action, or analyzing the situation in a dynamic environment. In contrast to ego-vehicle actions, other road users may be occluded or far away, posing 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 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 overrepresented scenarios based on occlusion rate and the 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

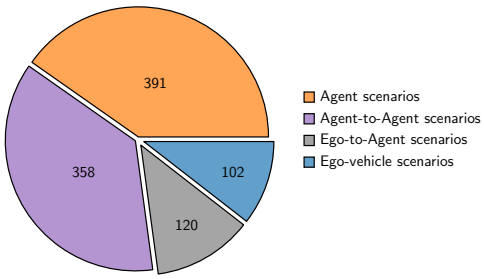


Figure 4: **Scenario distribution.** Number of scenarios per category.

questions offering at least five 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 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 the supplementary material.

Verification and correction. To ensure accuracy, the verification was conducted by three different individuals 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 needed to meet. The agreement on correctly mined scenarios was 85.6% (1017 agreements out of 1188) while there was a disagreement of 20.8% (247 out of 1188) on assigned negative scenarios, which results in a Fleiss' kappa of 0.87 and 0.97, respectively. This confirms consistent verifications across all experts. There were mainly differences in the perception of recognizable and unrecognizable agents, as well as 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 the supplementary material. For the final STSnu benchmark, we merged verified scenarios applying a majority voting and kept all assigned negative scenarios with full agreement over all reviewers. This conservative approach is designed to minimize ambiguity in answer choices and ensure high annotation quality.

4 Experiments

The evaluation on STSnu follows a simple protocol. For each scenario, we measure accuracy as the proportion of multiple-choice questions answered correctly. 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 (*i.e.*, ego-vehicle and agent trajectories). Therefore, we selected two open-source models, DeepSeek V3 [11] and Llama 3.2 [1], and one closed-source model, GPT-4o [45]. Second, for the evaluation of VLMs without fine-tuning, we use Qwen2.5-VL 7B [5], InternVL 2.5 1B [8], and InternVL 2.5 8B [8]. Third, as representative driving expert models, we evaluate Senna-VLM [28], OmniDrive [57], and DriveMM [21].

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 the supplementary material.

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 ego vehicle (ego-to-agent), 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 (58.93% vs. 47.79%) and ego-to-agent (75.15% vs. 61.83%) scenarios. GPT-4o [45], the advanced reasoning model, performs particularly well, reaching an average accuracy

Table 2: **Overall performance.** Performance comparison of LLMs (text-only), off-the-shelf VLMs (multi-frame camera images), and driving expert VLMs (multi-view videos). The evaluation is run five times using different randomly selected negative scenarios (*i.e.*, wrong answer choices). Accuracies are grouped by scenario categories. The best results are highlighted in bold.

	Ego	Ego-to-Agent	Agent	Agent-to-Agent	Average
Llama 3.2 [1]	24.77 ±1.7	20.28 ± 5.0	17.31 ±1.7	25.93 ±2.9	22.07 ±1.8
DeepSeek V3 [11]	48.39 ±9.8	59.07 ± 8.2	43.19 ±2.3	**45.18** ±2.7	48.96 ±1.9
GPT-4o [45]	**58.93** ±3.1	**75.15** ± 2.8	**44.03** ± 2.2	42.97 ±3.4	**55.27** ±2.0
InternVL 2.5 1B [8]	19.93 ±4.3	**46.68** ± 8.9	23.86 ±6.5	24.76 ±3.9	28.81 ±3.4
Qwen2.5-VL 7B [5]	37.37 ±3.6	42.36 ± 7.5	35.87 ±4.0	37.84 ±1.6	38.36 ±1.8
InternVL 2.5 8B [8]	**38.43** ±2.4	44.58 ± 8.9	**47.19 ±1.2**	**43.81 ±4.7**	**43.50** ±3.6
Senna-VLM [28]	8.81 ±1.9	44.44 ± 7.2	26.10 ±5.0	**31.45** ±3.1	27.70 ±2.7
OmniDrive [57]	24.40 ±5.6	42.97 ± 8.8	23.78 ±1.7	26.15 ±3.8	29.33 ±2.2
DriveMM [21]	**47.79** ±5.0	**61.83** ±11.2	**38.70** ±2.3	27.95 ±4.6	**44.07** ±2.8

of 55.27%. DriveMM [21] performs with 44.07%, on average, second best to GPT-4o [45]. It leads among all visual models, but only by a narrow margin ahead of the off-the-shelf InternVL 2.5 8B [8].

An interesting observation is the considerable gap between DriveMM [21] and the other expert models, particularly for the ego scenarios. While OmniDrive [57] projects multi-view image features into BEV, DriveMM directly processes multi-view videos. It is worth noting that the StreamPETR [56] encoder of OmniDrive [57] was initially designed for perception tasks, such as 3D object detection, where it is crucial 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 and qualitative results of all models in the supplementary material.

Number of multiple-choice options. To accommodate varying difficulty levels, our benchmark allows scaling the number of multiple-choice options. As shown in Fig. 5, increasing the number of options leads to a decrease in the accuracy of the DriveMM [21] model. This performance drop suggests that model predictions may also depend on the process of eliminating incorrect choices rather than a true understanding of the underlying scenario. Additionally, the increase in options results in greater variability in model performance, indicating increased task difficulty. We find that using five multiple-choice options offers an effective trade-off, which avoids both triviality and excessive complexity in the benchmark.

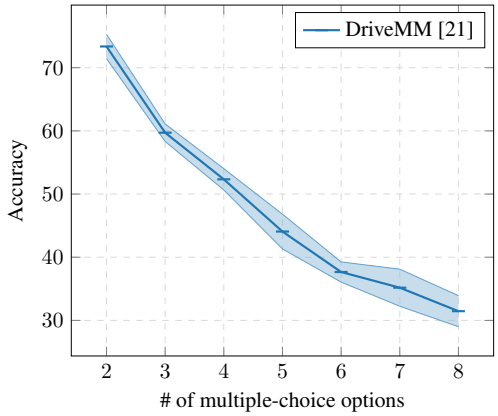


Figure 5: **Model performance over number of multiple-choice options.** Mean and variance of the DriveMM [21] accuracy across five independent evaluation runs, shown for increasing numbers of multiple-choice options (*e.g.*, two: A-B; three: A-B-C). In each evaluation, the order of the multiple-choice options was randomly permuted.

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 [6] 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 based on available data already pose significant challenges for driving expert models in terms of their spatio-temporal reasoning capabilities. Another bottleneck of automated 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 preselection 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.

Future work. With the growing interest in end-to-end autonomous driving and the corresponding expansion of the research community, an increasing number of high-quality datasets are becoming available. As perception models begin to reach saturation and deliver robust performance in standard driving conditions, research focus is shifting toward more challenging and rare situations. This trend opens up promising opportunities for extending STSBench to mine edge-case scenarios, such as accidents, wild animal crossings, or unexpected events, including cargo items (e.g., boxes, mats) falling from vehicles.

Beyond the inclusion of rare scenarios, STSBench can be extended along several additional dimensions. One direction is the modeling of finer-grained scenarios, incorporating risk attributes such as relative distances, agent velocities, or acceleration profiles to capture varying levels of criticality. Another extension involves multi-agent or event-chaining reasoning, where scenarios unfold across multiple entities and questions build upon one another. For example, a model might first identify the scenario as "overtaking" and then be asked: "Is this overtaking scenario dangerous?"

Once model performance begins to saturate on the current benchmark, which already poses a significant challenge to existing driving expert models, the benchmark can be extended to include a free-form answer track. This would challenge models to go beyond fixed-choice reasoning and offer richer explanations of driving situations, thereby providing a deeper assessment of their spatio-temporal understanding. In addition, more diversified question generation could be introduced to test the linguistic robustness of models and ensure that performance gains reflect true reasoning ability rather than sensitivity to phrasing.

Acknowledgments

The financial support by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged. We further acknowledge the EuroHPC Joint Undertaking for awarding us access to Leonardo at CINECA, Italy.

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] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

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

[39] 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.

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

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

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

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

[56] 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.

[57] 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.

[58] 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.

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

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Sec. 3 details the general benchmark extraction framework and the specific instantiation on NuScenes. Sec. 4 presents the evaluations and highlights the shortcomings of current models.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are explicitly discussed within Sec. 5.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Implementation details (incl. input formatting for all models) to support reproducibility are included in the supplementary material.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: As per the submission requirements of the NeurIPS datasets & benchmark track, both data and code are already openly accessible. The code and documentation allows reproducing all presented results.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The main manuscript includes all information necessary to understand the results, while (due to the page limit) the detailed parametrization of all evaluated models is included in the supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Sec. 4 presents the average results including standard deviation across 5 independent evaluation runs.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Implementation details are provided in the supplementary material.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All points of the NeurIPS code of ethics as well as our institution's ethics & scientific integrity guidelines have been respected.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader societal impacts are discussed within Sec. 5.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our framework and benchmark poses no high risk for misuse, and relevant limitations have been explicitly discussed.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Original works are referenced throughout the paper and framework, and licenses of all used assets have been respected.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Framework and the provided benchmark contain documentation to ensure their usability and reproducibility.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper involved no crowdsourcing or research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: No crowdsourcing or research with human subjects.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLM usage did not not impact the core methodology, scientific rigorousness, or originality of the research.