STRIDE-QA: Visual Question Answering Dataset for Spatiotemporal Reasoning in Urban Driving Scenes

Keishi Ishihara[1*], **Kento Sasaki**[1, 2*], **Tsubasa Takahashi**[1], **Daiki Shiono**[1, 3], **Yu Yamaguchi**[1]
[1]Turing Inc.
[2]University of Tsukuba
[3]Tohoku University
{keishi.ishihara, kento.sasaki}@turing-motors.com

Abstract

Vision-Language Models (VLMs) have been applied to autonomous driving to support decision-making in complex real-world scenarios. However, their training on static, web-sourced image-text pairs fundamentally limits the precise spatiotemporal reasoning required to understand and predict dynamic traffic scenes. We address this critical gap with STRIDE-QA, a large-scale visual question answering (VQA) dataset for physically grounded reasoning from an ego-centric perspective. Constructed from 100 hours of multi-sensor driving data in Tokyo, capturing diverse and challenging conditions, STRIDE-QA is the largest VQA dataset for spatiotemporal reasoning in urban driving, offering 16 M QA pairs over 270 K frames. Grounded by dense, automatically generated annotations including 3D bounding boxes, segmentation masks, and multi-object tracks, the dataset uniquely supports both object-centric and ego-centric reasoning through three novel QA tasks that require spatial localization and temporal prediction. Our benchmarks demonstrate that existing VLMs struggle significantly, with near-zero scores on prediction consistency. In contrast, VLMs fine-tuned on STRIDE-QA exhibit dramatic performance gains, achieving 55% success in spatial localization and 28% consistency in future motion prediction, compared to near-zero scores from general-purpose VLMs. Therefore, STRIDE-QA establishes a comprehensive foundation for developing more reliable VLMs for safety-critical autonomous systems.

Dataset — https://turingmotors.github.io/stride-qa/
Extended version — https://arxiv.org/abs/2508.10427

1 Introduction

Recent advances in Vision-Language Models (VLMs) have led to remarkable progress across multimodal tasks such as image captioning and visual question answering (Liu et al. 2023a; Bai et al. 2025). These models, trained on large-scale image-text datasets, exhibit strong semantic understanding and generalization capabilities. Motivated by this success, VLMs have been applied to Physical AI (NVIDIA et al. 2025), such as in robotics (Brohan et al. 2023; Black et al. 2024) and autonomous driving (Sima et al. 2024; Marcu et al. 2024).

These approaches underscore the promise of VLMs in achieving holistic scene understanding and high-level driving decision-making. However, this promise is fundamentally limited by the nature of their training data: most VLMs are trained on static, web-sourced image-text pairs and consequently lack the spatial reasoning capabilities essential for real-world applications (Fu et al. 2025). This limitation is particularly critical in autonomous driving, where the lack of appropriate training data remains a significant challenge.

To address this limitation, we introduce **STRIDE-QA** (**S**patio**T**emporal **R**easoning **I**n **D**riving Scenarios for **E**go-centric Visual **Q**uestion **A**nswering), a large-scale VQA dataset designed for fine-grained spatial and spatiotemporal reasoning in real-world driving scenes. An overview of the dataset and its automated annotation pipeline is shown in Figure 1. The dataset is constructed from over 100 hours of multi-sensor driving data collected in Tokyo, capturing diverse and challenging scenarios including traffic congestion, construction zones, and pedestrian-dense intersections. It contains over 16 M QA pairs generated through a fully modular and scalable annotation pipeline that integrates 3D object detection, multi-object tracking, and instance segmentation.

STRIDE-QA is specifically constructed to enable supervised training and evaluation on three core reasoning tasks:

- **Object-centric Spatial QA**: Assessing spatial relations between non-ego agents (vehicles, pedestrians, etc.).
- **Ego-centric Spatial QA**: Describing agents' distance, orientation, and size relative to the ego vehicle.
- **Ego-centric Spatiotemporal QA**: Predicting how agent–ego spatial relations evolve over time.

These tasks are designed to systematically measure the spatial and predictive reasoning capabilities essential for downstream planning and decision making in safety-critical urban environments. By grounding each QA pair in physically and temporally consistent annotations, STRIDE-QA provides a comprehensive foundation for training and benchmarking VLMs in real-world autonomous driving.

Contributions The main contributions of this paper are summarized as follows:

[*]These authors contributed equally.

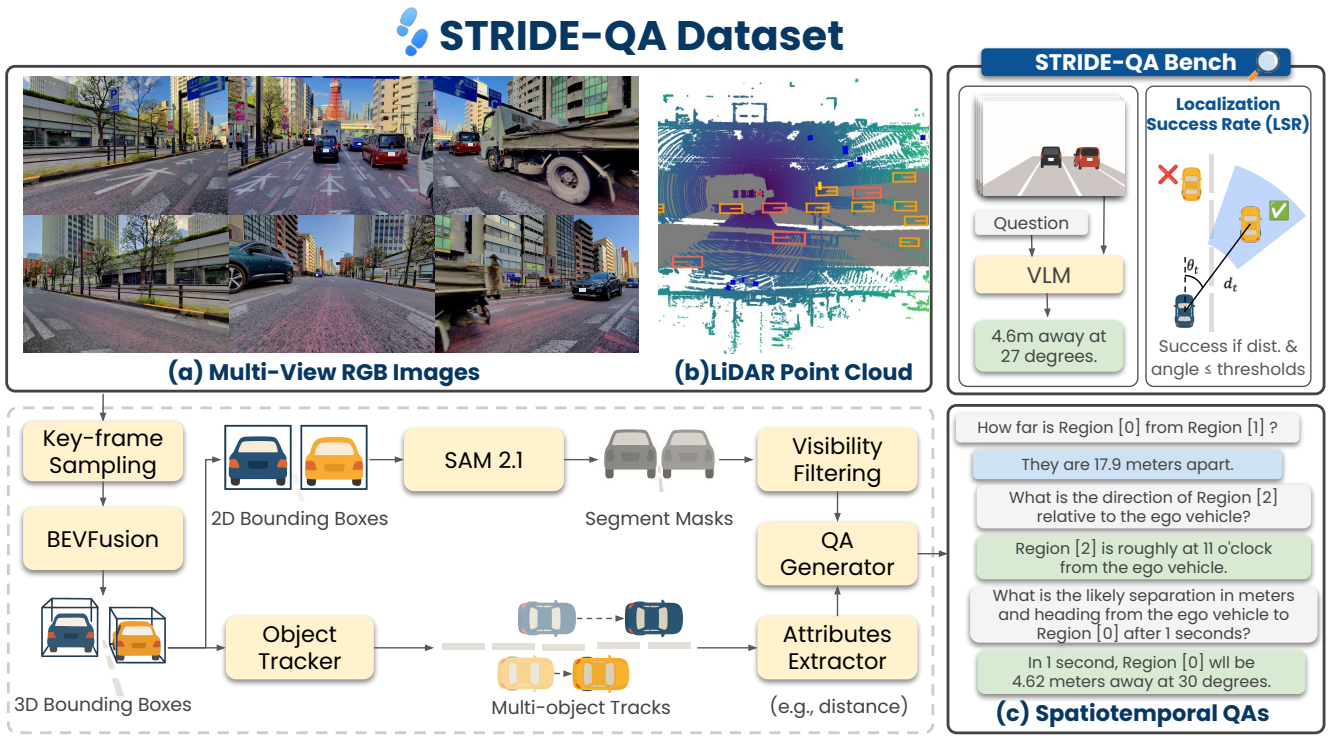

Figure 1: STRIDE-QA is a large-scale VQA dataset for spatiotemporal reasoning in autonomous driving, comprising 270 K frames and 16 M QA pairs from over 100 hours of urban driving in Tokyo. (a) It includes multi-view RGB images and (b) LiDAR point clouds, processed via a modular pipeline with 3D object detection, segmentation, tracking, and visibility filtering to produce spatially and temporally grounded annotations. (c) The annotations enable object-centric, ego-centric spatial, and spatiotemporal QA tasks, allowing structured evaluation of physically grounded reasoning over time.

- We define three novel ego-centric VQA tasks that jointly require spatial grounding and short-term predictive reasoning, addressing core challenges autonomous driving systems face in complex traffic scenes.
- We present STRIDE-QA, a large-scale dataset containing 16 M QA pairs densely annotated over 270 K video frames from urban driving, which enables supervised training of VLMs on fine-grained spatial and short-term temporal reasoning grounded in real-world traffic dynamics.
- We demonstrate that existing general-purpose VLMs struggle with spatiotemporal reasoning, while models fine-tuned on STRIDE-QA significantly outperform baselines. Our best model, STRIDE-Qwen2.5-VL-7B, achieves state-of-the-art performance, highlighting the effectiveness of our dataset for spatiotemporal understanding in driving scenes.

These contributions represent a key step toward integrating VLMs into real-world autonomous systems. By shifting from general-purpose vision-language understanding to physically grounded, ego-centric reasoning, STRIDE-QA bridges the gap between large-scale multimodal pretraining and the demands of Physical AI. It offers not only a benchmark for spatiotemporal VQA but also advances spatial understanding in dynamic scenes, promoting more trustworthy VLMs for autonomous driving.

2 Related Work

VLMs have demonstrated strong performance on multimodal tasks by training on 2D vision-language datasets such as VQA v2 (Goyal et al. 2017) and GQA (Ainslie et al. 2023). However, these datasets lack 3D spatial information, which limits their applicability to physically grounded reasoning in real-world environments.

To address this gap, spatially-aware VLMs such as SpatialVLM (Chen et al. 2024a) and Spatial-RGPT (Cheng et al. 2024) incorporate geometric annotations that enable metric reasoning in 3D. Nevertheless, these models remain confined to static, single-frame inputs and cannot capture the temporal dynamics critical for applications such as autonomous driving.

In response, several VQA datasets have been introduced for driving scenarios. For example, nuScenes-QA (Park et al. 2025) leverages multi-view videos to formulate spatial queries. ToD3Cap (Jin et al. 2024) and NuPrompt (Wu et al. 2025) focus on ego-centric spatial grounding, whereas Refer-KITTI (Wu et al. 2023b) and TUMTraffic-VideoQA (Zhou et al. 2025) explore video-based QA and referring expressions.

While these datasets represent significant progress toward incorporating spatial and temporal reasoning into vision-

Dataset	Data Source	Modality	# Video	# QA	Viewpoint		QA Type			
					Obj.	Ego	S-Q	S-N	ST-Q	ST-N
Spatial-VQA (Chen et al. 2024a)	Web images	RGB	—	200 M	✓		✓	✓		
Open Spatial Dataset (Cheng et al. 2024)	OpenImages	RGB	—	8.7 M	✓		✓	✓		
Refer-KITTI (Wu et al. 2023a)	KITTI	RGB + LiDAR	6 h	818		✓	✓		✓	
ToD3Cap (Jin et al. 2024)	nuScenes	RGB + LiDAR	5.5 h	468 K	✓	✓	✓			
nuScenes-QA (Qian et al. 2024)	nuScenes	RGB + LiDAR	5.5 h	460 K	✓	✓	✓		✓	
NuPrompt (Wu et al. 2025)	nuScenes	RGB + LiDAR	5.5 h	87.3 K		✓	✓		✓	
nuPlanQA (Park et al. 2025)	nuPlan	RGB + LiDAR	119 h	1 M	✓	✓	✓		✓	
TUMTraffic-VideoQA (Zhou et al. 2025)	AD	Self-collected	RGB	$\leq$33.3 h	87.3 K	✓	✓	✓		✓
STRIDE-QA (Ours)	Self-collected	RGB + LiDAR	100 h	16M	✓	✓	✓	✓		✓

Table 1: Comparison of STRIDE-QA with existing visual question answering datasets by data source, modality, scale, viewpoint, and QA types. S-Q, S-N, ST-Q, and ST-N denote Spatial Qualitative, Spatial Numerical, Spatiotemporal Qualitative, and Spatiotemporal Numerical. STRIDE-QA is the first dataset to provide large-scale, ego-centric spatiotemporal supervision.

language models, several challenges remain unaddressed. Notably, most existing benchmarks are designed to support either object-centric or ego-centric reasoning, but not both. This distinction is important because object-centric reasoning is crucial for capturing interactions among agents, whereas ego-centric reasoning enables spatial understanding relative to the ego vehicle. In addition, many datasets lack temporally aligned 3D annotations, essential for short-horizon predictive reasoning. Scalability also remains a challenge, as real-world driving scenarios involve high visual diversity and complexity. A detailed comparison of these datasets is provided in Table 1.

To fill this gap, we introduce STRIDE-QA. It comprises over 100 hours of synchronized LiDAR and multi-view RGB data, and defines three QA tasks: object-centric spatial, ego-centric spatial, and ego-centric spatiotemporal. These tasks collectively support fine-grained, predictive reasoning under complex traffic conditions.

3 Spatial and Spatiotemporal QA Definition

This section outlines key requirements for spatiotemporal QA in autonomous driving before introducing our dataset and benchmark. To support fine-grained reasoning in urban traffic scenarios, we define three VQA categories targeting distinct aspects of scene understanding: (1) relationships between surrounding agents, (2) ego-to-agent spatial relations, and (3) short-term prediction of agent dynamics.

Object-centric Spatial QA. This category targets spatial relationships between two surrounding agents based on a single current frame. It includes both qualitative questions (e.g., relative position or yes/no queries) and quantitative ones that require numerical answers (e.g., "1.53 m away" or "5 degrees"), as illustrated in Figure 2 (A).

Ego-centric Spatial QA. Understanding the scene from the ego-centric perspective is critical for autonomous driving. To support this, we design a category of questions that target the spatial relationship between the ego vehicle and a single surrounding agent, based on a single current frame. We construct both qualitative and quantitative questions involving distance, relative direction, and object size with respect to the ego vehicle, as illustrated in Figure 2 (B).

Ego-centric Spatiotemporal QA. This category extends Ego-centric Spatial QA by incorporating short-term temporal prediction. The task takes four context frames sampled at 2 Hz as input and aims to forecast the following physical quantities at future time horizons $t \in \{1, 2, 3\}$ s:

- **Distance** between the ego vehicle and the target agent
- **Heading angle** from the ego vehicle to the target agent
- **Velocity** of both the ego vehicle and the target agent

This task's QA examples are shown in Figure 2 (C).

4 STRIDE-QA Dataset

We introduce **STRIDE-QA** (**S**patio**T**emporal **R**easoning **i**n **D**riving Scenarios for **E**go-centric Visual **Q**uestion **A**nswering), a large-scale benchmark designed to advance ego-centric spatiotemporal reasoning in urban driving. STRIDE-QA consists of **270 K** frames, **268 K** unique object pairs, and **16 M** QA pairs.

Section 4.1 describes the data collection setup, and Section 4.2 details the automated annotation pipeline.

4.1 Driving Data Collection

Driving Areas. We collected over 100 hours of driving data in Tokyo, which is known for its challenging environment caused by traffic congestion, complex regulations such as no-parking zones and one-way streets, and the high density of pedestrians and cyclists. The city includes downtown districts, residential neighborhoods, and suburban areas.

Sensor Setup. We utilize a multi-purpose vehicle equipped with a sensor suite consisting of a 64-channel LiDAR and six cameras, which are synchronized and downsampled to 2 Hz. The cameras, each with a 60° field of view (FOV), offer 360° visual coverage with varied resolutions (front/back: 2880x1860 pixels; sides: 1920x1240 pixels). To accurately estimate the vehicle's state and position, we also record signals from an inertial measurement unit (IMU) and a real-time kinematic global navigation satellite system (RTK-GNSS) receiver. During data acquisition, all sensors and control area network (CAN) signals are recorded with

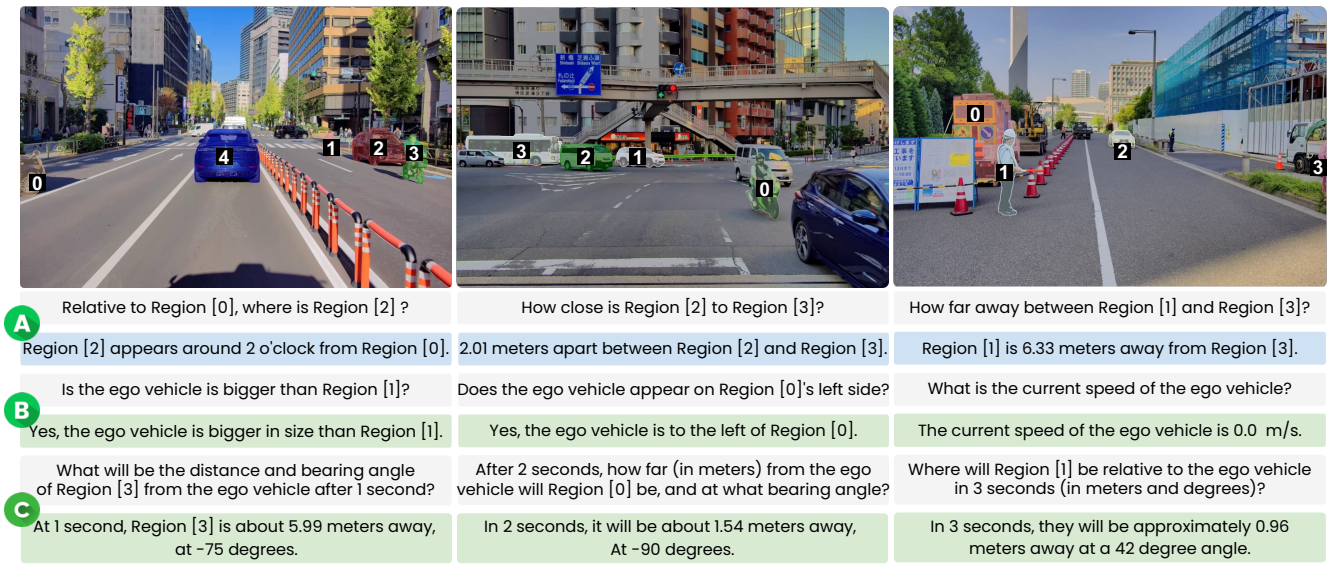

Figure 2: Example data from our STRIDE-QA dataset. From top to bottom, each QA pair corresponds to (A) Object-centric Spatial QA, (B) Ego-centric Spatial QA, and (C) Ego-centric Spatiotemporal QA.

synchronized timestamps. Following the nuScenes (Caesar et al. 2020), the recorded sequences are segmented into 20-second video clips. Further details about the sensor setup are provided in Appendix A.1.

4.2 Annotation Pipeline

We propose an automated annotation pipeline that processes sequences of synchronized multi-view RGB images and LiDAR point clouds to generate spatiotemporal question-answer pairs, each consisting of a question, an answer, and supporting visual evidence (e.g., bounding boxes or segmentation masks).

Our pipeline, illustrated in Figure 1, comprises seven components: keyframe sampling, 3D object detection, multi-object tracking, attributes extraction, semantic segmentation, visibility filtering, and question generation.

Keyframe Sampling. In our pipeline, a keyframe serves as the temporal anchor for each spatiotemporal question. We sample these at 1 Hz from the 2 to 17-second interval of each clip, which provides each keyframe with a temporal context window spanning from 2 seconds prior to 3 seconds after.

3D Object Detection. We adopt BEVFusion (Liu et al. 2023b), which fuses LiDAR and multi-camera images to perform high-precision 3D object detection. For each keyframe, we estimate the 3D position, orientation, and size of all visible objects. The set of object classes includes those from nuScenes as well as additional custom classes, totaling 45 categories.

Object Tracking. To capture temporal dynamics, we track the movement of each object over time. PubTracker (Yin, Zhou, and Krahenbuhl 2021) is selected for its simplicity and efficiency in point-based 3D object tracking, which enables fast and consistent ID assignment across frames without relying on appearance features. We associate 3D bounding boxes across consecutive frames and assign consistent instance IDs to each object. As a result, all frames, including context and future frames associated with each keyframe, are annotated with 3D bounding boxes, and objects can be consistently tracked over time.

Attributes Extractor To support spatiotemporal QA generation, we extract relevant attributes for each object in every frame. Based on the 3D bounding boxes and the pose of the ego vehicle, we compute the Euclidean distance and heading angle in the ego-centric coordinate system. In addition to these numerical values, we also generate discrete and qualitative spatial descriptions such as “left” or “1 o’clock” derived from these quantities.

Moreover, by leveraging the temporally consistent object tracks obtained during the tracking stage, we estimate the velocity of each object from the change in the center position of its bounding box over time. These attributes, including distance, heading, and velocity, are associated with each object in every frame and form the foundation for generating spatiotemporal QA pairs.

Semantic Segmentation. To obtain precise 2D segmentation masks for downstream visibility filtering and to provide fine-grained, pixel-level visual grounding for VLMs, we apply SAM 2.1 Large (Ravi et al. 2025). The model generates a class-agnostic mask within each object’s projected 3D bounding box. Each mask is then associated with the object’s consistent instance ID from the tracking stage.

Visibility Filtering. Since bounding boxes and segmentation masks are automatically generated, it is important to filter out unreliable objects to ensure annotation quality. We apply the following three filtering rules: (1) the Intersection-over-Union (IoU) between bounding boxes must be greater than 0.3, (2) the coverage rate must exceed 0.8, and (3) the IoU score predicted by SAM 2.1 Large must also exceed 0.8. By applying these rules, we effectively eliminate noisy detections and retain only high-quality annotations.

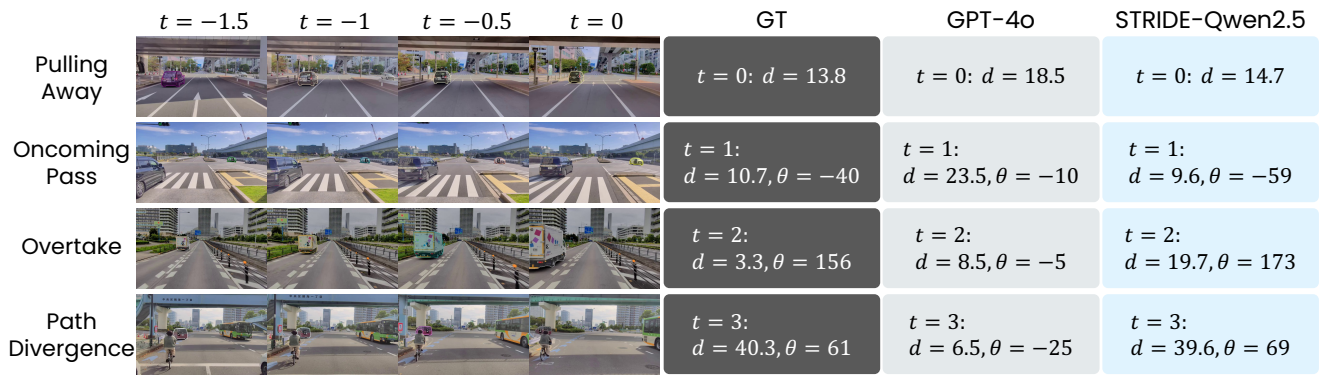

Figure 3: Qualitative results on STRIDE-QA Bench. Across four driving scenarios, our fine-tuned STRIDE-Qwen2.5-VL-7B (labeled as STRIDE-Qwen2.5) consistently delivers more accurate distance and angle estimates than GPT-4o. Note that model responses are abbreviated to highlight key numerical predictions (t: time [s], d: distance [m], θ: angle [°]).

Question Generator. Finally, we generate template-based QA pairs for each keyframe based on the previously extracted object attributes, such as distance, heading, and velocity. Each object is referred to as `Region [X]` in the question text, where the number `X` corresponds to the visual identifier shown in the image. These QA pairs are aligned with the 2D bounding boxes and segmentation masks from the current, context, and future frames. Each QA pair is associated with a keyframe t_k, and refers to a target object or object pair within a temporal window $[t_k - 2, t_k + 3]$. The spatial grounding s consists of 2D bounding boxes and segmentation masks across the relevant frames t, aligned via consistent instance IDs. As illustrated in Figure 2, this results in a dataset composed of paired image, segmentation, and QA examples.

These stages automatically yield spatiotemporal QA pairs from synchronized multi-camera RGB and LiDAR data, ensuring temporally aligned, instance-consistent annotations for VLM training and evaluation; annotation quality details are in Appendix A.

5 Experiments

We evaluate whether STRIDE-QA improves VLMs' spatiotemporal reasoning in traffic scenes by benchmarking multiple models on spatial and spatiotemporal QA tasks.

5.1 Experimental Setup

Fine-tuned Models. We fine-tune two open-source VLMs, Qwen2.5-VL-7B-Instruct and Cosmos-Reason1-7B, on the training split of STRIDE-QA using only the front-camera RGB sequence ($t \in \{-1.5, -1.0, -0.5, 0\}$ s). Side and rear camera images and LiDAR are omitted to match the evaluation setup detailed in the next subsection. We refer to the resulting models as STRIDE-Qwen2.5-VL-7B and STRIDE-Cosmos-Reason1-7B.

Baseline Models. We evaluate our fine-tuned models against a range of VLMs. To estimate the upper bound of general-purpose VLMs, we include proprietary GPT-4o and GPT-4o mini (OpenAI et al. 2024). We further compare with the open-source models InternVL2.5-8B (Chen et al. 2024b) and Qwen2.5-VL-7B-Instruct (Bai et al. 2025), the spatially enhanced SpatialRGPT-8B (Cheng et al. 2024), and autonomous driving-specific VLMs, Senna-VLM (Jiang et al. 2024) and Cosmos-Reason1-7B (NVIDIA et al. 2025). Details of the training process are provided in Appendix B.

Spatial Benchmark. We evaluate spatial reasoning on the outdoor split of the SpatialRGPT-Bench (Cheng et al. 2024). The details of the SpatialRGPT-Bench are provided in Appendix D.

Spatiotemporal Benchmark. Beyond spatial reasoning, we introduce **STRIDE-QA Bench** to evaluate spatiotemporal reasoning capabilities. We employ three primary types of metrics, which are briefly described in Section 5.2 and detailed in Appendix C.

5.2 STRIDE-QA Bench

STRIDE-QA Bench is an evaluation suite for spatiotemporal reasoning in urban driving scenes. Beyond spatial QA benchmarks, we assess the performance of spatiotemporal QA tasks defined in Section 3. We report four metrics: Localization Success Rate (LSR), Mean Localization Success Rate (MLSR), Temporal Localization Consistency (TLC), and diagnostic per-dimension success rates (SR).

Evaluation Setup. The model is given a sequence of four RGB frames from a front-facing onboard camera with a 60° FOV, captured at $t \in \{-1.5, -1.0, -0.5, 0\}$ s. In this sequence, the model observes a single *target agent* from one of six classes (*car*, *large_vehicle*, *bus*, *pedestrian*, *motorcycle*, or *bicycle*), identified by its segmentation mask in all four frames to provide sufficient context. The model is tasked with predicting the following quantities at $t \in \{0, 1, 2, 3\}$ s: the target agent's *distance*, *velocity*, and *heading angle*, along with the ego vehicle's *velocity*. The agent's heading angle is defined in the ego vehicle's frame of reference, where 0° is forward and positive angles are counterclockwise on the range $(-180°, 180°]$. This definition is explicitly provided in the prompt for fair evaluation. Notably, the task includes predicting agents that may move out of the camera's FOV at $t > 0$.

Model	Original ↑		Obj. Spatial ↑		Ego Spatial ↑	
	Qual.	Quant.	Qual.	Quant.	Qual.	Quant.
GPT-4o	**80.5**	32.5	**68.1**	39.4	55.7	27.7
GPT-4o mini	54.7	30.6	56.3	32.1	57.1	44.4
Intern-VL 2.5 8B	64.1	18.8	48.4	26.6	36.3	29.1
Qwen2.5-VL 7B-Instruct	67.2	24.4	64.0	12.8	47.1	29.3
SpatialRGPT-VILA-1.5-8B	75.0	**46.9**	58.4	42.2	25.3	16.9
Senna-VLM	18.0	0.63	9.14	4.59	5.88	2.03
Cosmos-Reason1-7B	53.9	30.0	55.2	21.1	33.2	20.3
STRIDE-Qwen2.5-VL-7B	69.5	37.5	61.1	**61.5**	77.9	**70.3**
STRIDE-Cosmos-Reason1-7B	71.1	30.0	62.2	58.7	**79.9**	68.9

Table 2: SpatialRGPT-Bench results. Evaluation of qualitative (Qual.) and quantitative (Quant.) spatial reasoning performance on the original SpatialRGPT dataset, as well as our object-centric and ego-centric extensions.

Model	LSR ↑				MLSR ↑	TLC ↑
	0s	1s	2s	3s		
GPT-4o	18.1	6.6	6.1	7.6	9.6	0.7
GPT-4o mini	4.6	2.0	0.7	0.7	2.0	0.0
InternVL2.5-8B	2.4	1.0	1.7	0.7	1.5	0.0
Qwen2.5-VL-7B-Instruct	1.0	3.4	4.4	1.0	2.4	0.0
SpatialRGPT-VILA-1.5-8B	0.5	0.2	0.2	0.0	0.2	0.0
Senna-VLM	1.0	0.0	0.2	0.0	0.3	0.0
Cosmos-Reason1-7B	1.5	3.2	2.0	1.5	2.0	0.0
STRIDE-Qwen2.5-VL-7B	96.3	**46.2**	**38.4**	**38.9**	**55.0**	**28.4**
STRIDE-Cosmos-Reason1-7B	**96.8**	43.5	37.4	36.2	53.5	25.4

Table 3: STRIDE-QA Bench results. our fine-tuned models achieve the best performance across LSR@t, MSLR, and TLC metrics.

Evaluation Dataset. The evaluation dataset is built from held-out recording dates of our STRIDE-QA corpus to ensure train-test separation. We define a *scene group* as a sequence centered on a single target agent observed across four context frames ($t \in -1.5, -1.0, -0.5, 0$s) and evaluated across four future timesteps ($t \in 0, 1, 2, 3$s). Each scene group includes 13 QA pairs, totaling 5,317 QA pairs across 409 scene groups. To focus on challenging spatiotemporal reasoning, we filter the data to exclude static and repetitive scenes, retaining only those with dynamic interactions. These scene groups are categorized into six dynamic scenarios: *Oncoming Pass*, *Maintain State*, *Overtake*, *Path Divergence*, *Pulling Away From Ego*, and *Minor Relations*. We define an out-of-view (OOV) event where the target object completely exits the front camera's FOV in any future frame ($t \in 1, 2, 3$s). These scenarios exhibit a clear divide in their OOV likelihood; for instance, scenarios like *Oncoming Pass* and *Overtake* consistently involve OOV events, whereas others like *Maintain State* rarely do. Detailed statistics for all scenarios are provided in Appendix C.

Localization Success Rate (LSR). LSR is a binary metric that evaluates whether a model's spatial prediction is simultaneously accurate in both distance and direction. Here, $\hat{d}_t$ and $\hat{\theta}_t$ denote model predictions, while d_t^* and θ_t^* are the corresponding ground-truth values. A prediction is successful only if the estimated distance and orientation, expressed in polar coordinates $\hat{\mathbf{p}}_t = (\hat{d}_t, \hat{\theta}_t)$, satisfy

$$|\hat{d}_t - d_t^*| < 0.25\, d_t^* \quad \text{and} \quad |\hat{\theta}_t - \theta_t^*| < 10^\circ.$$

We adopt a $\pm 25\%$ distance margin following prior work (Cheng et al. 2024) and set the $\pm 10^\circ$ heading margin so that at 10 m the lateral deviation equals a standard 3.5 m lane width.

Mean Localization Success Rate (MLSR). MLSR averages the per-frame LSR over a sequence, providing a softer measure of temporal stability:

$$\text{MLSR} = \frac{1}{|G|} \sum_{g \in G} \frac{1}{T+1} \sum_{t=0}^{T} s_{g,t},$$

where $\mathbf{s}_g = [s_{g,0}, \ldots, s_{g,T}]$ denotes the LSR success bits for scene $g \in G$.

Temporal Localization Consistency (TLC). TLC is a strict metric that measures the proportion of sequences where the agent is successfully localized across all four timesteps, i.e., the target is never lost:

$$\text{TLC} = \frac{1}{|G|} \sum_{g \in G} \mathbf{1}\big[s_{g,0} \wedge s_{g,1} \wedge s_{g,2} \wedge s_{g,3}\big].$$

where $\mathbf{s}_g = [s_{g,0}, \ldots, s_{g,T}]$ is the sequence of LSR success bits for a scene group g with $T = 3$.

Per-dimension Success Rates. Single-axis success rates for distance, heading angle, and velocities are computed for diagnostic analysis. Full metric definitions and detailed results are provided in Appendix C.

5.3 Results on SpatialRGPT-Bench

Table 2 summarizes the results. STRIDE-Qwen2.5-VL-7B and STRIDE-Cosmos-Reason1-7B substantially outperform their base models across all splits. In the quantitative Object-centric Spatial QA, STRIDE-Qwen2.5-VL-7B rises from 12.8 to 61.5 (+48.7 pt, ×4.8) and STRIDE-Cosmos-Reason1-7B from 21.1 to 58.7 (+37.6 pt, ×2.8). Similar gains appear in the quantitative Ego-centric split, where scores improve from 29.3 to 70.3 (+41.0 pt) and from 20.3 to 68.9 (+48.6 pt) for the two models, respectively. These results indicate that training on the STRIDE-QA dataset boosts fine-grained spatial reasoning on external benchmarks. Although GPT-4o and SpatialRGPT-8B achieve the highest accuracy on the original split, their performance drops markedly on the Object-centric subset. This gap suggests that generic large models do not automatically transfer to camera-centric setups, possibly due to viewpoint shifts and annotation noise. A detailed ablation is provided in Appendix D.

5.4 Results on STRIDE-QA Bench

As presented in Table 3, all baseline models, including powerful general-purpose VLMs, exhibit poor performance on the spatiotemporal benchmark defined in Section 5.2. Their Localization Success Rate (LSR) scores are low, and they achieve near-zero results on the multi-frame (MLSR) and

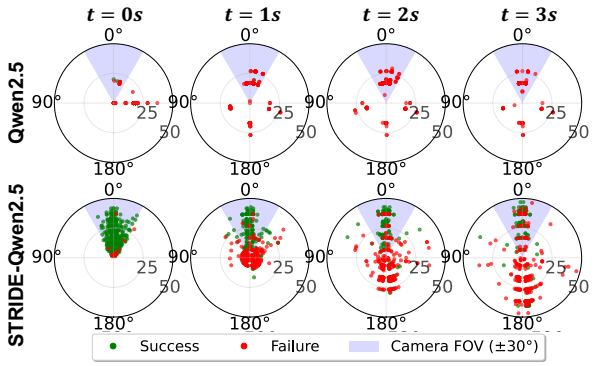

Figure 4: Comparison of localization (angle=heading, radius=ego distance). Finetuning (bottom) boosts performance vs. base (top): dense, plausible forecasts and near-perfect $t = 0\,s$ localization. Green/red dots indicate LSR success/-failure; the blue wedge is the camera $\pm 30°$ FOV.

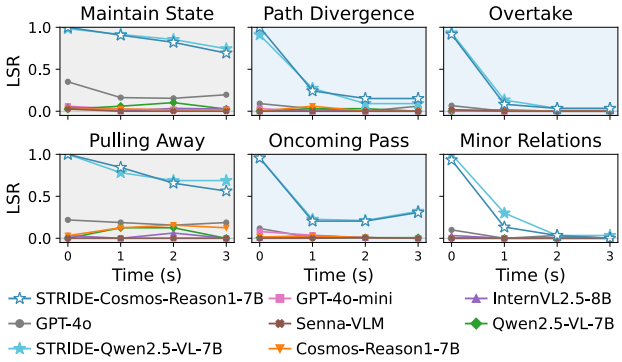

Figure 5: LSR trends across dynamic scenarios, grouped by out-of-view (OOV) likelihood. The sharp performance drop in scenarios with a high OOV likelihood (blue background) compared to in-view scenarios (grey background) highlights that OOV prediction is the primary challenge.

temporal consistency (TLC) metrics. This indicates a fundamental inability of existing models to perform reliable spatiotemporal reasoning in complex driving scenarios.

In contrast, models fine-tuned on STRIDE-QA demonstrate dramatic performance gains. Our top-performing model, STRIDE-Qwen2.5-VL-7B achieves an $LSR_{t=0}$ of 96.3% (96× baseline), demonstrating precise spatial understanding. This extends to future horizons with an $LSR_{t=3}$ of 38.9% (39× baseline). Furthermore, it obtains an MLSR of 55.0 and a TLC of 28.4, confirming that our dataset effectively teaches consistent reasoning across viewpoints and time (see Figure 3 for qualitative examples).

However, the results also highlight the task's difficulty. While a TLC score of 28.4 improves from 0, it indicates that achieving short-term consistent reasoning remains challenging. This suggests that fine-tuning on our dataset helps, but may be insufficient to capture the complexity of real-world dynamic prediction.

In conclusion, our work makes two key contributions. First, we show that fine-tuning on STRIDE-QA closes a critical gap in spatiotemporal reasoning capabilities of VLMs. Second, by quantifying remaining challenges in short-term consistency, our benchmark guides future research toward building robust, safety-critical autonomous systems.

6 Analysis

Beyond performance metrics, we conduct in-depth analysis to better understand the strengths, limitations, and generalization behaviors of VLMs trained with STRIDE-QA.

6.1 Qualitative Analysis of Prediction Patterns

A qualitative analysis of prediction patterns in Figure 4 reveals fundamental differences in model behavior. The baseline VLM (top row) exhibits a consistent failure mode: its predictions are not only sparse but also systematically biased, repeating similar incorrect guesses regardless of the visual input. This suggests the model defaults to a simplistic, memorized behavior instead of grounding its reasoning in the visual context. In contrast, our fine-tuned model (bottom row) is responsive, generating dense predictions across the 360-degree space and accurately localizing objects in the present frame ($t = 0s$). The baseline's failure to produce context-aware, continuous predictions demonstrates a lack of temporal consistency, which is a fundamental prerequisite for motion planning. This prerequisite is often overlooked in prior work connecting VLMs to planning modules (Sima et al. 2024; Tian et al. 2024), a gap our framework is designed to address by providing a framework to explicitly train and evaluate this skill.

6.2 Error Analysis on Out-of-View Prediction

To understand the root cause of the temporal degradation identified in the previous section, we analyze our model's performance across the different dynamic scenarios defined in our benchmark (Figure 5). A clear pattern emerges: in scenarios where the target agent is likely to remain within the camera's FOV, such as *Maintain State*, the LSR declines gracefully. Conversely, in scenarios where the agent tends to exit the FOV, such as *Overtake*, *Oncoming Pass*, and *Path Divergence*, the LSR degrades much more sharply. This stark contrast strongly suggests that the primary failure mode for long-term prediction is the model's inability to reason about OOV object trajectories. This finding exposes a key limitation of relying on single-camera input for forecasting and highlights that integrating multi-camera information is a critical direction to build robust autonomous driving.

7 Conclusion

We introduced STRIDE-QA, a large-scale VQA dataset addressing spatiotemporal reasoning gaps in autonomous-driving VLMs. Fine-tuning on it boosted performance: our best model reached 55.0% MLSR and 28.4 TLC, dramatically surpassing otherwise near-zero baselines. STRIDE-QA establishes a foundation for robust, physically grounded vision-language understanding in autonomous systems.

Acknowledgments

This paper is based on results obtained from GENIAC (Generative AI Accelerator Challenge, a project to strengthen Japan's generative AI development capabilities), a project (JPNP20017) implemented by the Ministry of Economy, Trade and Industry (METI) and the New Energy and Industrial Technology Development Organization (NEDO).

References

Ainslie, J.; Lee-Thorp, J.; De Jong, M.; Zemlyanskiy, Y.; Lebrón, F.; and Sanghai, S. 2023. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints. arXiv:2305.13245.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Baruch, G.; Chen, Z.; Dehghan, A.; Dimry, T.; Feigin, Y.; Fu, P.; Gebauer, T.; Joffe, B.; Kurz, D.; Schwartz, A.; et al. 2021. Arkitscenes: A Diverse Real-World Dataset for 3D Indoor Scene Understanding Using Mobile RGB-D Data. https://arxiv.org/abs/2111.08897. arXiv:2111.08897.

Black, K.; Brown, N.; Driess, D.; Esmail, A.; Equi, M.; Finn, C.; Fusai, N.; Groom, L.; Hausman, K.; Ichter, B.; Jakubczak, S.; Jones, T.; Ke, L.; Levine, S.; Li-Bell, A.; Mothukuri, M.; Nair, S.; Pertsch, K.; Shi, L. X.; Tanner, J.; Vuong, Q.; Walling, A.; Wang, H.; and Zhilinsky, U. 2024. π_0: A Vision-Language-Action Flow Model for General Robot Control. arXiv:2410.24164.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Chen, X.; Choromanski, K.; Ding, T.; Driess, D.; Dubey, A.; Finn, C.; Florence, P.; Fu, C.; Arenas, M. G.; Gopalakrishnan, K.; Han, K.; Hausman, K.; Herzog, A.; Hsu, J.; Ichter, B.; Irpan, A.; Joshi, N.; Julian, R.; Kalashnikov, D.; Kuang, Y.; Leal, I.; Lee, L.; Lee, T.-W. E.; Levine, S.; Lu, Y.; Michalewski, H.; Mordatch, I.; Pertsch, K.; Rao, K.; Reymann, K.; Ryoo, M.; Salazar, G.; Sanketi, P.; Sermanet, P.; Singh, J.; Singh, A.; Soricut, R.; Tran, H.; Vanhoucke, V.; Vuong, Q.; Wahid, A.; Welker, S.; Wohlhart, P.; Wu, J.; Xia, F.; Xiao, T.; Xu, P.; Xu, S.; Yu, T.; and Zitkovich, B. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. arXiv:2307.15818.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11621–11631.

Chen, B.; Xu, Z.; Kirmani, S.; Ichter, B.; Sadigh, D.; Guibas, L.; and Xia, F. 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14455–14465.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24185–24198.

Cheng, A.-C.; Yin, H.; Fu, Y.; Guo, Q.; Yang, R.; Kautz, J.; Wang, X.; and Liu, S. 2024. SpatialRGPT: Grounded Spatial Reasoning in Vision-Language Models. In *NeurIPS*.

Fu, X.; Hu, Y.; Li, B.; Feng, Y.; Wang, H.; Lin, X.; Roth, D.; Smith, N. A.; Ma, W.-C.; and Krishna, R. 2025. BLINK: Multimodal Large Language Models Can See but Not Perceive. In Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Computer Vision – ECCV 2024*, 148–166. Cham: Springer Nature Switzerland. ISBN 978-3-031-73337-6.

Geiger, A.; Lenz, P.; and Urtasun, R. 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 3354–3361.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6904–6913.

Gupta, V. 2023. Dashcam Anonymizer. https://github.com/varungupta31/dashcam_anonymizer. Accessed: 2025-08-31.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Jiang, B.; Chen, S.; Liao, B.; Zhang, X.; Yin, W.; Zhang, Q.; Huang, C.; Liu, W.; and Wang, X. 2024. Senna: Bridging Large Vision-Language Models and End-to-End Autonomous Driving. arXiv:2410.22313.

Jin, B.; Zheng, Y.; Li, P.; Li, W.; Zheng, Y.; Hu, S.; Liu, X.; Zhu, J.; Yan, Z.; Sun, H.; Zhan, K.; Jia, P.; Long, X.; Chen, Y.; and Zhao, H. 2024. TOD3Cap: Towards 3D Dense Captioning in Outdoor Scenes. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29 – October 4, 2024, Proceedings, Part XVIII*, 367–384. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-031-72648-4.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual Instruction Tuning. In *NeurIPS*.

Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D.; and Han, S. 2023b. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Marcu, A.-M.; Chen, L.; Hünermann, J.; Karnsund, A.; Hanotte, B.; Chidananda, P.; Nair, S.; Badrinarayanan, V.; Kendall, A.; Shotton, J.; Arani, E.; and Sinavski, O. 2024. LingoQA: Visual Question Answering for Autonomous Driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 252–269.

NVIDIA; :; Azzolini, A.; Brandon, H.; Chattopadhyay, P.; Chen, H.; Chu, J.; Cui, Y.; Diamond, J.; Ding, Y.; Ferroni,

F.; Govindaraju, R.; Gu, J.; Gururani, S.; Hanafi, I. E.; Hao, Z.; Huffman, J.; Jin, J.; Johnson, B.; Khan, R.; Kurian, G.; Lantz, E.; Lee, N.; Li, Z.; Li, X.; Lin, T.-Y.; Lin, Y.-C.; Liu, M.-Y.; Luo, A.; Mathau, A.; Ni, Y.; Pavao, L.; Ping, W.; Romero, D. W.; Smelyanskiy, M.; Song, S.; Tchapmi, L.; Wang, A. Z.; Wang, B.; Wang, H.; Wei, F.; Xu, J.; Xu, Y.; Yang, X.; Yang, Z.; Zeng, X.; and Zhang, Z. 2025. Cosmos-Reason1: From Physical Common Sense To Embodied Reasoning. arXiv:2503.15558.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; Currier, J.; Dai, Y.; Decareaux, C.; Degry, T.; Deutsch, N.; Deville, D.; Dhar, A.; Dohan, D.; Dowling, S.; Dunning, S.; Ecoffet, A.; Eleti, A.; Eloundou, T.; Farhi, D.; Fedus, L.; Felix, N.; Fishman, S. P.; Forte, J.; Fulford, I.; Gao, L.; Georges, E.; Gibson, C.; Goel, V.; Gogineni, T.; Goh, G.; Gontijo-Lopes, R.; Gordon, J.; Grafstein, M.; Gray, S.; Greene, R.; Gross, J.; Gu, S. S.; Guo, Y.; Hallacy, C.; Han, J.; Harris, J.; He, Y.; Heaton, M.; Heidecke, J.; Hesse, C.; Hickey, A.; Hickey, W.; Hoeschele, P.; Houghton, B.; Hsu, K.; Hu, S.; Hu, X.; Huizinga, J.; Jain, S.; Jain, S.; Jang, J.; Jiang, A.; Jiang, R.; Jin, H.; Jin, D.; Jomoto, S.; Jonn, B.; Jun, H.; Kaftan, T.; Łukasz Kaiser; Kamali, A.; Kanitscheider, I.; Keskar, N. S.; Khan, T.; Kilpatrick, L.; Kim, J. W.; Kim, C.; Kim, Y.; Kirchner, J. H.; Kiros, J.; Knight, M.; Kokotajlo, D.; Łukasz Kondraciuk; Kondrich, A.; Konstantinidis, A.; Kosic, K.; Krueger, G.; Kuo, V.; Lampe, M.; Lan, I.; Lee, T.; Leike, J.; Leung, J.; Levy, D.; Li, C. M.; Lim, R.; Lin, M.; Lin, S.; Litwin, M.; Lopez, T.; Lowe, R.; Lue, P.; Makanju, A.; Malfacini, K.; Manning, S.; Markov, T.; Markovski, Y.; Martin, B.; Mayer, K.; Mayne, A.; McGrew, B.; McKinney, S. M.; McLeavey, C.; McMillan, P.; McNeil, J.; Medina, D.; Mehta, A.; Menick, J.; Metz, L.; Mishchenko, A.; Mishkin, P.; Monaco, V.; Morikawa, E.; Mossing, D.; Mu, T.; Murati, M.; Murk, O.; Mély, D.; Nair, A.; Nakano, R.; Nayak, R.; Neelakantan, A.; Ngo, R.; Noh, H.; Ouyang, L.; O'Keefe, C.; Pachocki, J.; Paino, A.; Palermo, J.; Pantuliano, A.; Parascandolo, G.; Parish, J.; Parparita, E.; Passos, A.; Pavlov, M.; Peng, A.; Perelman, A.; de Avila Belbute Peres, F.; Petrov, M.; de Oliveira Pinto, H. P.; Michael; Pokorny; Pokrass, M.; Pong, V. H.; Powell, T.; Power, A.; Power, B.; Proehl, E.; Puri, R.; Radford, A.; Rae, J.; Ramesh, A.; Raymond, C.; Real, F.; Rimbach, K.; Ross, C.; Rotsted, B.; Roussez, H.; Ryder, N.; Saltarelli, M.; Sanders, T.; Santurkar, S.; Sastry, G.; Schmidt, H.; Schnurr, D.; Schulman, J.; Selsam, D.; Sheppard, K.; Sherbakov, T.; Shieh, J.; Shoker, S.; Shyam, P.; Sidor, S.; Sigler, E.; Simens, M.; Sitkin, J.; Slama, K.; Sohl, I.; Sokolowsky, B.; Song, Y.; Staudacher, N.; Such, F. P.; Summers, N.; Sutskever, I.; Tang, J.; Tezak, N.; Thompson, M. B.; Tillet, P.; Tootoonchian, A.; Tseng, E.; Tuggle, P.; Turley, N.; Tworek, J.; Uribe, J. F. C.; Vallone, A.; Vijayvergiya, A.; Voss, C.; Wainwright, C.; Wang, J. J.; Wang, A.; Wang, B.; Ward, J.; Wei, J.; Weinmann, C.; Welihinda, A.; Welinder, P.; Weng, J.; Weng, L.; Wiethoff, M.; Willner, D.; Winter, C.; Wolrich, S.; Wong, H.; Workman, L.; Wu, S.; Wu, J.; Wu, M.; Xiao, K.; Xu, T.; Yoo, S.; Yu, K.; Yuan, Q.; Zaremba, W.; Zellers, R.; Zhang, C.; Zhang, M.; Zhao, S.; Zheng, T.; Zhuang, J.; Zhuk, W.; and Zoph, B. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Park, S.-Y.; Cui, C.; Ma, Y.; Moradipari, A.; Gupta, R.; Han, K.; and Wang, Z. 2025. NuPlanQA: A Large-Scale Dataset and Benchmark for Multi-View Driving Scene Understanding in Multi-Modal Large Language Models. arXiv:2503.12772.

Qian, T.; Chen, J.; Zhuo, L.; Jiao, Y.; and Jiang, Y.-G. 2024. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 4542–4550.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollar, P.; and Feichtenhofer, C. 2025. SAM 2: Segment Anything in Images and Videos. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

Roberts, M.; Ramapuram, J.; Ranjan, A.; Kumar, A.; Bautista, M. A.; Paczan, N.; Webb, R.; and Susskind, J. M. 2021. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10912–10922.

Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Beißwenger, J.; Luo, P.; Geiger, A.; and Li, H. 2024. DriveLM: Driving with Graph Visual Question Answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 256–274.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 567–576. Los Alamitos, CA, USA: IEEE Computer Society.

Tian, X.; Gu, J.; Li, B.; Liu, Y.; Wang, Y.; Zhao, Z.; Zhan, K.; Jia, P.; Lang, X.; and Zhao, H. 2024. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. In *8th Annual Conference on Robot Learning (CoRL)*.

Wu, D.; Han, W.; Liu, Y.; Wang, T.; Xu, C.-Z.; Zhang, X.; and Shen, J. 2025. Language Prompt for Autonomous Driving. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(8): 8359–8367.

Wu, D.; Han, W.; Wang, T.; Dong, X.; Zhang, X.; and Shen, J. 2023a. Referring Multi-Object Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14633–14642.

Wu, D.; Han, W.; Wang, T.; Dong, X.; Zhang, X.; and Shen, J. 2023b. Referring Multi-Object Tracking. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 14633–14642.

Yang, J.; Zhang, H.; Li, F.; Zou, X.; Li, C.; and Gao, J. 2023. Set-of-mark Prompting Unleashes Extraordinary Visual Grounding in GPT-4V. https://arxiv.org/abs/2310.11441. arXiv:2310.11441.

Yin, T.; Zhou, X.; and Krahenbuhl, P. 2021. Center-Based 3D Object Detection and Tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11784–11793.

Zhou, X.; Larintzakis, K.; Guo, H.; Zimmer, W.; Liu, M.; Cao, H.; Zhang, J.; Lakshminarasimhan, V.; Strand, L.; and Knoll, A. 2025. TUMTraf VideoQA: Dataset and Benchmark for Unified Spatio-Temporal Video Understanding in Traffic Scenes. In *Forty-second International Conference on Machine Learning*.