Spatial-aware Vision Language Model for Autonomous Driving

Supplementary Material

S1. Abstract

This supplementary material provides additional details and analyses of our work. Further information about our proposed SA-QA dataset is presented in Sec. S2, and a detailed comparison with NuScenes-SpatialQA is provided in Sec. S3. Additional details on the 3D grounding benchmark are included in Sec. S4, and limitations and future work are discussed in Sec. S5.

S2. More Details on SA-QA Dataset

This dataset is built on top of nuScenes [1] dataset and enriched using the ground-truth annotations provided by both nuScenes and OpenLane [2] to generate question-answering (QA) pairs. As SA-QA is designed for training, we use only the training split of the nuScenes dataset. The specific QA formats and the step-by-step generation procedure are summarized in Tab. S1 and Tab. S2. Note that before generating any QA pairs, we convert all annotations and LiDAR point clouds into the ego coordinate system.

S3. Comparison with NuScenes-SpatialQA

NuScenes-SpatialQA [3] is a concurrent study that shares certain similarities with our proposed SA-QA dataset. While both datasets focus on the spatial reasoning of Vision-Language Models (VLMs), SA-QA diverges by prioritizing instruction tuning, explicit metric grounding, and cross-modal alignment.

Training vs. Evaluation Focus. NuScenes-SpatialQA is constructed exclusively on the nuScenes validation split (150 scenes). Its primary purpose is to serve as a zero-shot benchmark to evaluate existing general-purpose VLMs. In contrast, SA-QA is designed as a massive-scale instruction-tuning dataset constructed on the training split (850 scenes). This scale allows LVLDrive to learn complex spatial relationships rather than merely being tested on them.

Enhanced Cross-Modality Interaction. To promote robust alignment across modalities, SA-QA introduces specific task designs targeting Text-Vision-LiDAR integration. We implement modality masking (Tab. S1, SP-04), which masks the target in the image to encourage the model to retrieve geometric information directly from the LiDAR data. Furthermore, we utilize visual cues (Tab. S1, SP-03; Tab. S2, SR-03), where prompts explicitly reference arrows drawn on the image. This design drives the model to bind language with joint image–LiDAR features, enabling richer 2D–3D spatial reasoning.

Explicit 3D Grounding vs. Relative Depth. While NuScenes-SpatialQA utilizes 3D annotations to generate answers, its questions are largely limited to relative 3D distances or topological relationships (*e.g.* "Is object A closer than object B?"), omitting explicit 3D locations. This allows models to rely on approximate depth cues without mastering metric space. Conversely, SA-QA dataset requires explicit 3D grounding, asking the model to output precise coordinates (x, y) and absolute dimensions (Tab. S1, SP-02). This forces the model to internalize a true metric understanding of the 3D environment.

Global Perception vs. Intra-View Limitations. A critical limitation of NuScenes-SpatialQA is that questions are typically restricted to the same camera view. This relies on local visual comparisons and omits a holistic perception of the surrounding environment. SA-QA explicitly challenges this by constructing cross-view reasoning tasks (Tab. S2, SR-03) where target objects may appear in disparate sensors (*e.g.* Front Camera vs. Back-Right Camera). To answer these correctly, the model cannot rely on a single 2D image but must fuse information into a unified global coordinate system.

Automated Efficiency vs. LLM Latency. NuScenes-SpatialQA relies on heavy Large Language Models to generate dense captions and formulate questions, a process that introduces significant computational latency and cost. In contrast, SA-QA is fully automated and rule-based. By deriving QA pairs directly from ground-truth annotations, our generation process is computationally negligible. This high efficiency supports dynamic augmentation, allowing us to apply random sampling strategies on the fly—specifically randomizing target objects, lane segments, and temporal intervals for future prediction.

S4. More Details on 3D Grounding Benchmark

Throughout model development, we observed that planning-only evaluation is insufficient and fails to capture a model's understanding of the spatial distribution of surrounding objects. To better assess this capability, we construct a grounding benchmark using the ground-truth annotations of the nuScenes [1] validation set.

Distance-Based Object Sampling. Traversing all annotated objects would lead to prohibitive inference time, so we sample target objects based on distance. For each frame, we compute the distance from the bottom center of every annotated 3D bounding box to the ego vehicle and sort all objects by distance. The objects are then grouped into four ranges. After experimenting with interval sizes of 15 m, 20 m, and

Table S1. **Generation logic for spatial perception (SP) tasks in the SA-QA dataset.**

ID	Prompt Template	Input Data Transformation	Generation Logic	Answer Generation
SP-01	"For a potential future position at (x, y), is it in a drivable area?"	None.	1. Randomly sample a query point in the front region of the ego-vehicle ($5 \leq X \leq 20$, $-5 \leq Y \leq 5$). 2. Generate a drivable mask in BEV by buffering lane centerlines with a 1.75m margin. 3. Check point inclusion against the mask.	**Binary:** "Yes" if inside; "No" otherwise.
SP-02	"Identify the object in $<CAM, x_{min}, y_{min}, x_{max}, y_{max}>$ and describe its 3D information."	None.	1. Project 3D annotations to the 2D planes of all 6 cameras. 2. Compute 2D bounding boxes $[x_{min}, y_{min}, x_{max}, y_{max}]$ clamped to image dims, filtering out candidates that are invisible or too small. 3. Sample an object and format the answer. 4. Format the prompt with the target object's coordinates.	**Text:** "The object is a <category> in the <CAM>, location: (x, y), length: $<l>$, width: $<w>$, height: $<h>$, angles in degree: $<yaw>$." (Rounded to 0.1).
SP-03	"Identify the object cued by the **arrow** and describe its 3D information."	Draw an arrow on the image.	1–3. Follow steps 1–3 of SP-02 to filter candidates and sample one object. 4. Draw a visual arrow on the image pointing to the center of the target's 2D mask. 5. Construct the prompt referencing the arrow cue.	**Text:** Same format as SP-02.
SP-04	"Identify the object in the **masked region** and describe its 3D information."	Mask a region of the image.	1–3. Follow steps 1–3 of SP-02 to filter candidates and sample one object. 4. Apply a mask to the target's 2D bounding box region (forcing LiDAR reliance). 5. Construct the prompt referencing the masked region.	**Text:** Same format as SP-02.

Table S2. **Generation logic for spatial reasoning (SR) tasks in the SA-QA dataset.**

ID	Prompt Template	Input Data Transformation	Generation Logic	Answer Generation
SR-01	"What objects are on the lane defined by points $(x_1, y_1), (x_2, y_2), (x_3, y_3)$?"	None.	1. Randomly select a lane centerline from the OpenLane [2] map annotations. 2. Aggregate the associated objects located on the selected lane centerline from the OpenLane annotation set [2]. 3. Format the prompt and answer.	**List:** "The object is a <category>, location...".
SR-02	"What is the nearest object in the <DIRECTION> direction?"	None.	1. Define 4 spatial sectors (*e.g.* Front-Left, Back-Left) relative to the current ego-vehicle heading. 2. Filter objects located within the target area. 3. Calculate Euclidean distances for all candidates and sort in ascending order. 4. Format the prompt and answer.	**Text:** Description of the object with index 0 (minimum distance), using the format from SP-02.
SR-03	"Please determine the metric distance (in meters) separating the two indicated objects."	Draw arrows on the images.	1. Select two distinct visible objects (O_A, O_B), potentially across different camera views. 2. Draw visual arrows pointing to O_A and O_B in their respective images (following SP-03). 3. Compute the L2 norm between their 3D centroids: $\|C_A - C_B\|_2$.	**Scalar:** "D." (The value is rounded to 0.1 meters).
SR-04	"What is the future position of the object at (x, y) after T second?"	None.	1. Randomly sample one object that possesses a future trajectory within the nuScenes annotation set [1]. 2. Randomly select a future position and time interval for the sampled object, and subsequently structure the input prompt and the corresponding target answer.	**Coordinate:** "(x_{fut}, y_{fut})."

Table S3. **Distribution of object counts across distance ranges in our grounding benchmark.**

0-15m	15-30m	30-45m	45m-Inf	Overall
5,304	5,684	5,367	3,903	20,258

25 m, we found that a 15 m step yields the most balanced distribution across groups, and thus adopt 15 m as the interval. Finally, for each frame, we randomly sample one object from each distance range to form our grounding benchmark. The number of objects in each range is shown in Tab. S3.

Question and Answer Generation. For each selected object, we format its 3D bounding box parameters using the following answer template: The object is a <class> in the <direction>, location: (X, Y), length: L, width: W, height: H, angles in degrees: θ. Meanwhile, we project the eight corners of the 3D box onto the image plane and compute the corresponding 2D bounding box $(x_{min}, y_{min}, x_{max}, y_{max})$, which is incorporated into the question. If any of these coordinates fall outside the image bounds, they are clipped to the valid image range.

Answer Parsing and Evaluation. Because similar QA patterns are also used during training, the model typically produces answers that closely follow the desired template. This allows us to reliably parse a BEV bounding box (X, Y, L, W, θ) from the predicted answer and compute the BEV mIoU. Concretely, we first compute the IoU for each individual object. We then average the IoUs within each of the four distance ranges, and finally report the mean of these four group-wise IoUs as the overall grounding metric.

S5. Limitations and Future Work.

Despite the benefits of our gradual fusion strategy, LVLDrive is still constrained by the limited availability of large-scale, naturally paired text–LiDAR data. In an ideal setting, one would pretrain a unified model on large-scale aligned language–LiDAR samples, enabling the representation space of 3D geometry and linguistic concepts to be co-optimized from scratch. The construction of such datasets remains extremely challenging, and closing this gap is an important direction for future work.

References

[1] 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. 1, 2

[2] Li Chen, Chonghao Sima, Yang Li, Zehan Zheng, Jiajie Xu, Xiangwei Geng, Hongyang Li, Conghui He, Jianping Shi, Yu Qiao, and Junchi Yan. Persformer: 3d lane detection via perspective transformer and the openlane benchmark. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2

[3] 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. In *ICCVw*, 2025. 1