Supplementary Material for "NuScenes-SpatialQA: A Spatial Understanding and Reasoning Benchmark for Vision-Language Models in Autonomous Driving"

A. Details about raw data

NuScenes [1] dataset consists of 1,000 diverse urban driving scenes collected in Boston and Singapore, each lasting 20 seconds and recorded at 2 Hz. The dataset provides fully annotated 3D object detection and tracking data, featuring synchronized multi-sensor recordings from six cameras, a 32-beam LiDAR, five radars, and additional vehicle state information such as GPS and IMU.

It provides 3D annotations for 23 object categories, including vehicles, pedestrians, traffic cones, and barriers. Each annotated object is represented by a 3D bounding box with attributes such as position, size, orientation, and visibility level. The annotations are available at 2 Hz across 1,000 urban driving scenes.

B. Implementation Details for Auto Captioning

B.1. Implementation Details for 3D-to-2D Bounding Box Projection

We utilize the official nuScenes development kit (*nuscenes-devkit*) for projecting 3D LiDAR bounding boxes onto the 2D image plane. Specifically, it transforms 3D bounding boxes from the LiDAR coordinate system to the camera coordinate system and applies the intrinsic camera matrix for projection. The projected points are post-processed to determine the 2D bounding box coordinates. The specific code used is from the official nuScenes repository: https://github.com/nutonomy/nuscenes-devkit/blob/master/python-sdk/nuscenes/scripts/export_2d_annotations_as_json.py

B.2. Implementation Details for Visibility and Resolution Filter

To ensure the quality of 2D bounding boxes used in our evaluation, we apply a filtering process based on object size and visibility. Specifically, we remove bounding boxes with a **width or height** ≤ 40 **pixels**, as these objects are too small to provide meaningful visual information. Additionally, we exclude objects with a **visibility token** < 4, indicating that they are not fully visible in the camera view. According to the nuScenes definition, the visibility token is categorized into four levels: 1 (invisible), 2 (occluded), 3 (partially visible), and 4 (fully visible). By retaining only fully visible objects ($\texttt{visibility} = 4$), we eliminate ambiguous or heavily occluded instances, ensuring a cleaner and more reliable dataset for evaluation.

B.3. Implementation Details for Grouping Object Instances to Optimize Cropping

In a given scene, the same object may appear across multiple keyframes. However, if we extract and store every instance of an object from each keyframe and use it as input for the VLM, this would lead to excessive memory consumption and significantly slow down the caption generation process. Additionally, since each keyframe contains multiple objects, processing every frame individually would create a large number of redundant inputs, many of which may not contribute meaningful new information. Furthermore, not all cropped images from every frame will result in high-quality captions due to variations in object visibility, occlusion, and resolution. Therefore, a more efficient strategy is needed to select representative frames for each object while maintaining high caption quality.

To efficiently select the best frames for generating high-quality captions, we group bounding boxes by object instance within each scene. Instead of processing every appearance of an object across all keyframes, we aggregate all its bounding boxes throughout the scene based on its `instance_token`. This allows us to analyze the object's occurrences holistically and choose the most representative frames for caption generation.

We select **Llama-3.2-11b-Instruct** as the VLM for caption generation. For each selected object instance, we extract its cropped image and feed it into the VLM with the following prompt:

Prompt for Caption Generation

```
Provide a short noun phrase captioning
{category_name} in the center of the
image, such as 'black sedan with red
logo' or 'man in a blue t-shirt and
jeans'.  The response must be a phrase
only.  Do NOT include full sentences
or extra descriptions.
```

This prompt encourages the model to generate concise yet discriminative captions.

C. Implementation Details for Auto Generation of 3D Scene Graphs

C.1. Node Structure

The node structure in our scene graph is represented using the following JSON format:

Node Structure: JSON Representation


```
{
    "node_id": "
        a721d524937f4a228fa6aac3296fb3bc",
    "attributes": {
        "category_name": "human.pedestrian.
            adult",
        "translation": {
            "x": 286.706,
            "y": 926.831,
            "z": 1.176
        },
        "size": {
            "length": 1.095,
            "width": 0.695,
            "height": 1.78
        },
        "caption": "the man in tan t-shirt and
             jeans"
    }
}
```


Here, `node_id` corresponds to the instance token, which uniquely identifies an object across different frames. The `attributes` field contains essential properties of the object, including its `category_name`, `translation` (3D coordinate), `size` (physical dimension), and a `caption` generated by the auto-captioning process.

C.2. Edge Structure

The edge structure in our scene graph is represented using the following JSON format:

Edge Structure: JSON Representation


```
{
    "edge_id": "296fb3bc_b43a1a15",
    "from": "
        a721d524937f4a228fa6aac3296fb3bc",
    "to": "069
        a7d8902d14560b1890064b43a1a15",
    "spatial_distance": 1.25,
    "longitudinal_offset": 0.65,
    "lateral_offset": 1.06,
    "relative_bearing_angle": -121.66
}
```


Each edge in the graph encodes spatial relationships between objects. The attributes are defined as follows: `edge_id` is a unique identifier for the edge, while `from` and `to` represent the unique IDs of the connected nodes (objects). `spatial_distance` denotes the Euclidean distance between the two objects in meters. `longitudinal_offset` refers to the displacement along the ego vehicle's heading direction, whereas `lateral_offset` indicates the perpendicular displacement relative to the ego vehicle's heading. Finally, `relative_bearing_angle` represents the angle (in degrees) from the `from` node to the `to` node in the ego vehicle's reference frame.

D. Details for QA Template

QA Template: Qualitative

- **Q:** Is {*object_1*} above {*object_2*}?
  A: yes / no
- **Q:** Is {*object_1*} below {*object_2*}?
  A: yes / no
- **Q:** Is {*object_1*} to the left of {*object_2*}?
  A: yes / no
- **Q:** Is {*object_1*} to the right of {*object_2*}?
  A: yes / no
- **Q:** Is {*object_1*} in front of {*object_2*}?
  A: yes / no
- **Q:** Is {*object_1*} behind {*object_2*}?
  A: yes / no
- **Q:** Is {*object_1*} larger than {*object_2*}?
  A: yes / no
- **Q:** Is {*object_1*} smaller than {*object_2*}?
  A: yes / no

- **Q:** Is {*object_1*} longer than {*object_2*} in length?
  A: yes / no
- **Q:** Is {*object_1*} shorter than {*object_2*} in length?
  A: yes / no
- **Q:** Is {*object_1*} taller than {*object_2*} in height?
  A: yes / no
- **Q:** Is {*object_1*} shorter than {*object_2*} in height?
  A: yes / no
- **Q:** Is {*object_1*} wider than {*object_2*}?
  A: yes / no
- **Q:** Is {*object_1*} thinner than {*object_2*}?
  A: yes / no

QA Template: Quantitative

- **Q:** In this image captured by {*camera*}, what is the distance between {*object_1*} and {*object_2*}?
  A: (numeric value)
- **Q:** In this image captured by {*camera*}, what is the distance between the ego vehicle and {*object*}?
  A: (numeric value)
- **Q:** In this image captured by {*camera*}, what is the longitudinal offset between the ego vehicle and {*object*}?
  A: (numeric value)
- **Q:** In this image captured by {*camera*}, what is the longitudinal offset between {*object_1*} and {*object_2*}?
  A: (numeric value)
- **Q:** In this image captured by {*camera*}, what is the lateral offset between the ego vehicle and {*object*}?
  A: (numeric value)
- **Q:** In this image captured by {*camera*}, what is the lateral offset between {*object_1*} and {*object_2*}?
  A: (numeric value)
- **Q:** What is the length of {*object*} in the image?
  A: (numeric value)
- **Q:** What is the width of {*object*} in the image?
  A: (numeric value)
- **Q:** What is the height of {*object*} in the image?
  A: (numeric value)
- **Q:** What is the relative bearing angle of {*object_2*} with respect to {*object_1*}?
  A: (numeric value)

D.1. Implementation Details for Generating Captions with VLM

QA Template: Direct Reasoning

- **Question:** From the given options, which object is the closest to {*object*}?
  (a) {*object_1*}
  (b) {*object_2*}
  (c) {*object_3*}
  (d) {*object_4*}
  Answer: a / b / c / d
- **Question:** This image is captured by one of the onboard cameras mounted on a vehicle. From the given options, which object is the closest to the ego vehicle?
  (a) {*object_1*}
  (b) {*object_2*}
  (c) {*object_3*}
  (d) {*object_4*}
  Answer: a / b / c / d
- **Question:** From the given options, which object is the largest in terms of overall size?
  (a) {*object_1*}
  (b) {*object_2*}
  (c) {*object_3*}
  (d) {*object_4*}
  Answer: a / b / c / d
- **Question:** Which object is closer to {*object_c*} in the image?
  (a) {*object_1*}
  (b) {*object_2*}
  Answer: a / b
- **Question:** This image is captured by one of the onboard cameras mounted on a vehicle. Which object is closer to the ego vehicle?
  (a) {*object_1*}
  (b) {*object_2*}
  Answer: a / b
- **Question:** Are there any pedestrians or vehicles within 5 meters of {*object*} in the image?
  Answer: yes / no
- **Question:** This image is captured by one of the onboard cameras mounted on a vehicle. Are there any

Models	LLaVA-v1.6	Llama-3.2	blip2	Qwen2.5-VL	Deepseek-vl2	SpatialRGPT
Backbone	Mistral-7B	Llama 3.2	Flan-T5-XL	Qwen2.5	MoE Transformer	pre-trained OpenAI CLIP-L
Parameter Size	7B	11B	3B	7B	3B	8B

Table 1. Baseline

pedestrians or vehicles within 10 meters of the ego vehicle?
Answer: yes / no

- **Question:** Are there any vehicles in the image with a width greater than 2 meters?
Answer: yes / no

QA Template: Situational Reasoning

- **Question:** This image is captured by one of the onboard cameras mounted on a vehicle. In autonomous driving, it is crucial to detect potential safety risks, especially when pedestrians are too close to vehicles. If a pedestrian is within 10 meters of a vehicle, it may indicate a potential hazard that requires caution. Given this, does the ego vehicle have a potential safety risk due to nearby pedestrians?
Answer: yes / no

- **Question:** Assume the distance between {*object_1*} and {*object_2*} is decreasing at 2 meters per second. Will they collide within 5 seconds?
Answer: yes / no

- **Question:** This image is captured by one of the onboard cameras mounted on a vehicle. Assume the ego vehicle is moving forward while all other objects remain stationary. Will there be a moment when {*object_1*} occludes {*object_2*}, causing {*object_2*} to become invisible from the ego vehicle's perspective?
Answer: yes / no

- **Question:** Assume there is a bridge ahead with a maximum clearance height of 2 meters. Any vehicle taller than this cannot safely pass under. Given this assumption, is there any vehicle in the current scene unable to pass under the bridge?
Answer: yes / no

- **Question:** Assume there is a parking spot measuring {*spot_length*} meters in length and {*spot_width*} meters in width. Considering that a vehicle needs at least {*clearance*} meters of clearance on both the front/back and left/right sides, can {*object*} in the image fit into this parking spot?
Answer: yes / no

- **Question:** Given the current distance between {*object_1*} and {*object_2*}, can a vehicle with a width of {*vehicle_width*} meters safely pass between them?
Answer: yes / no

E. Details for Baselines

Please refer to Table 1 for the backbone and parameter size of baseline VLMs.

F. Details for CoT

Prompt for CoT Reasoning

"You are given a question about spatial relationships in an autonomous driving scene. Think step by step before answering. First, analyze the spatial arrangement of objects based on the given context. Then, determine the correct answer based on your reasoning. Finally, provide your answer in the following format:
Reasoning: (Step-by-step explanation)
Answer: (Yes/No) / (A/ B) / (A/ B/ C/ D) *(according to question type)*
Question: {*question*}"

G. Broader Impact and Ethics Statement

G.1. Broader Impact Statement

NuScenes-SpatialQA provides a benchmark to evaluate the spatial reasoning capabilities of VLMs. Accurate spatial understanding is crucial for AI applications in autonomous driving, robotic navigation, and general visual perception. By systematically assessing VLMs' ability to interpret spatial relationships, our work helps identify limitations and guide improvements in AI-driven spatial reasoning.

G.2. Ethics Statement

Our research emphasizes fairness, transparency, and reliability. The benchmark is built on publicly available data while ensuring privacy and unbiased evaluation. We acknowledge the challenges of spatial reasoning in AI and advocate for responsible model development to minimize errors and unintended biases in real-world applications.

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