Inter-3D VQA: A Roadside Multimodal Benchmark for 3D Spatiotemporally Grounded Visual Question Answering

Shaozu Ding **Linan Song** **Dajiang Suo***
The Polytechnic School, Arizona State University
{sding32, lsong124, dsuo2}@asu.edu

Abstract

Recent advances in visual question answering (VQA) and multimodal large language models (MLLMs) have enabled natural-language reasoning over traffic scenes. However, existing benchmarks are largely built from ego-vehicle views or 2D roadside videos, limiting their ability to evaluate 3D-grounded reasoning over real-world distances, trajectories, infrastructure topology, and safety-critical interactions. We introduce Inter-3D VQA, a large-scale roadside multimodal benchmark for 3D spatiotemporally grounded VQA at intersections. Built from synchronized point clouds and multi-view images, Inter-3D VQA contains 407K QA pairs covering lane-level positions, object relationships, motion patterns, and near-miss-oriented interaction reasoning. We further propose Inter-Geo, an MLLM baseline that integrates object- and scene-level aligned LiDAR representations, and Inter-Metrics, a unified evaluation framework for textual consistency, numerical accuracy, and semantic correctness. Experiments show that Inter-Geo outperforms image-based VLMs, especially on grounded spatial and temporal reasoning tasks. Our benchmark and codes are available at https://github.com/ASU-Suo-Lab/Inter-3D-VQA.

1 Introduction

Recent advances in multimodal large language models (MLLMs) (Liu et al., 2023) have shifted traffic scene understanding from task-specific perception (Lang et al., 2019) toward language-guided reasoning (Qian et al., 2024). Visual question answering (VQA) (Antol et al., 2015) enables users and downstream systems to query complex traffic scenes through natural language, supporting driving-oriented reasoning (Deruyttere et al., 2019; Xu et al., 2024) such as identifying relevant road users and explaining potential risks. Beyond onboard systems, VQA also offers a promising interface for infrastructure-side traffic monitoring and control, where operators may ask about crosswalk occupancy, lane blockage, pedestrian distance to conflict points, or safety-critical interactions.

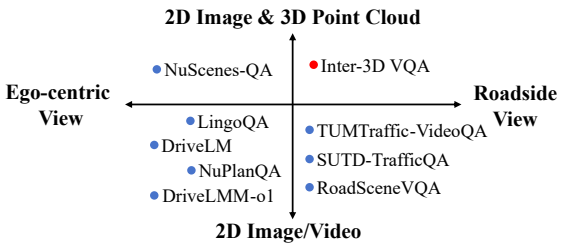


Figure 1: Inter-3D VQA is the first roadside multimodal benchmark for 3D grounded VQA.

However, existing traffic VQA benchmarks are still limited for 3D-grounded reasoning at intersection scale. As shown in Fig. 1, ego-centric benchmarks (Sima et al., 2024; Park et al., 2025) rely on onboard sensors and suffer from occlusion and local viewpoints, while roadside benchmarks (Zhou et al., 2025; Guan et al., 2026) offer infrastructure-side perspectives but mainly use single-view images or 2D videos. As a result, they lack explicit 3D grounding for evaluating models' understanding of real-world distances, trajectories, lane-level positions, and spatial relations among road users and infrastructure.

This limitation is critical for traffic monitoring and safety applications. As illustrated in Fig. 2, a system may need to localize a pedestrian in 3D space (Q1), determine whether the pedestrian remains within a crosswalk (Q2), and predict the trajectory of an approaching left-turning vehicle (Q3) to assess whether their future paths could result in a near-miss event (Q4), where road users come dangerously close without collision and signal potential safety risks before a crash occurs. Such reasoning requires spatiotemporal grounding in physical 3D space beyond image-level semantics. Due to the

*Corresponding author.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 34661–34686
October 24-29, 2026 ©2026 Association for Computational Linguistics

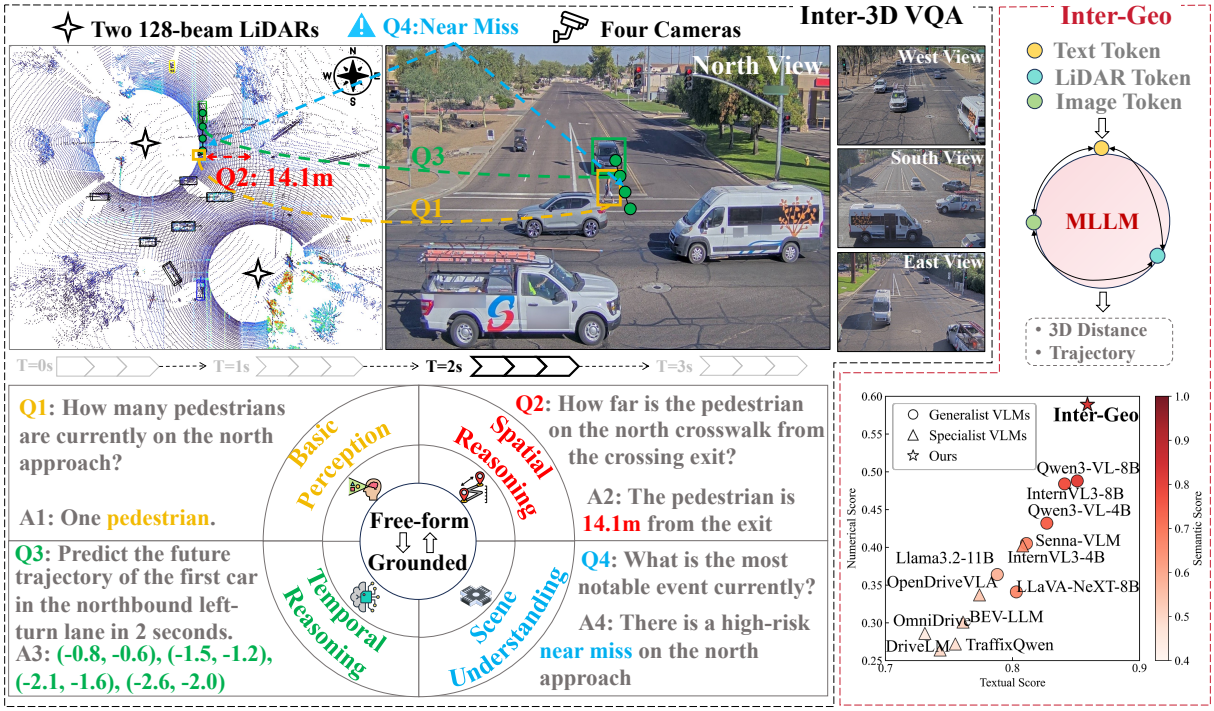


Figure 2: **Inter-3D VQA** is a large-scale roadside multimodal benchmark, aligning multi-view images with global point clouds to support diverse tasks from 3D perspective. Our MLLM **Inter-Geo** leverages object- and scene-level aligned LiDAR representations, achieving state-of-the-art performance on 3D grounded reasoning tasks.

lack of such data and benchmarks, the community still lacks a systematic way to develop and evaluate MLLMs for bird's-eye-view traffic understanding and 3D grounded VQA at intersections.

To address these challenges, we introduce Inter-3D VQA, a large-scale roadside multimodal benchmark built on synchronized global point clouds and multi-view images, as shown in Fig. 2. It explicitly integrates 3D spatial information and infrastructure topology into VQA tasks. We further propose Inter-Geo, an MLLM that aligns object- and scene-level LiDAR representations via Dual Querying Net (DQNet) and injects them into the decoder through a LiDAR Decoder Adapter (LDA). We also introduce Inter-Metrics, a unified evaluation framework for textual consistency, numerical accuracy, and semantic correctness. The main contributions of this work are summarized as follows:

- We present **Inter-3D VQA**, a large-scale multimodal dataset for intersection understanding. It contains 407K QA pairs built from multi-view images, global point clouds, and infrastructure topology, enabling 3D-aware reasoning in complex roadside scenes.
- We propose a comprehensive benchmark for 3D roadside VQA, covering two QA formats and four task categories. We further introduce **Inter-Metrics**, a unified evaluation protocol that jointly measures textual consistency, numerical accuracy, and semantic correctness.
- We propose **Inter-Geo**, a multimodal large language model that incorporates object- and scene-level aligned LiDAR representations. Extensive experiments demonstrate its advantages over image-based VLMs, especially in 3D spatiotemporal reasoning.

2 Related Works

Existing VQA research in transportation spans autonomous driving, urban scenes, and roadside understanding, with substantial differences in sensing perspectives, tasks, and data modalities. Table 1 summarizes representative VQA benchmarks across these scenarios in terms of task design, construction method, QA type, modality, and scale.

2.1 Autonomous Driving VQA

Existing autonomous driving VQA benchmarks are predominantly built on ego-centric views, focusing on driving-oriented reasoning. NuScenes-QA (Qian et al., 2024), LingoQA (Marcu et al., 2024) and DriveLM (Sima et al., 2024) establish

Dataset	Domain	Main Tasks	QA Generation	QA Type	Modality	QAs
NuScenes-QA (2024)	Driving	3D VQA	Temp.	Free-form	Image + PC	460k
LingoQA (2024)	Driving	VQA	Man. + LLM	Free-form	Image	419k
DriveLM (2024)	Driving	GVQA + T.P.	Man. + Temp.	Grounded	Image	443k
DriveLMM-o1 (2025)	Driving	M.C.	Man. + LLM	Grounded	Image	18k
NuPlanQA (2025)	Driving	M.C.	Man. + LLM	Free-form	Image	1M
SURDS (2025)	Driving	VQA	Temp. + LLM	Free-form	Image	50.3k
STSBench (2025)	Driving	M.C.	Temp.	Free-form	Image	971
Impromptu VLA (2025)	Driving	T.P.	Man. + LLM	Free-form	Image	80k
AutoDrive-QA (2025)	Driving	M.C.	Man. + LLM	Grounded	Image	15.4k
STRIDE-QA (2026)	Driving	VQA	Temp.	Grounded	Video	16M
V2X-QA (2026)	V2X	M.C.	Man. + LLM	Free-form	Image	33k
City-3DQA (2024)	City	3D VQA	Temp. + LLM + Man.	Free-form	Point Cloud	450k
Open3D-VQA (2025)	City	3D VQA	Temp. + LLM + Man.	Free-form	Image	73.3k
SUTD-TrafficQA (2021)	Roadside	VQA	Man.	Free-form	Video	62k
TUMTraffic-VideoQA (2025)	Roadside	VQA + M.C.	Temp. + LLM	Grounded	Video	87.3k
RoadSceneVQA (2026)	Roadside	VQA	Man. + LLM	Free-form	Image	34.7k
Inter-3D VQA (ours)	Roadside	3D VQA + T.P.	Temp. + LLM + Man.	Both	Image + PC	407k

Table 1: Comprehensive comparison of representative VQA benchmarks in autonomous driving, urban, and roadside scenarios. Temp.: Template, Man.: Manual, M.C.: Multiple Choice, T.P.: Trajectory Prediction, PC: Point Cloud.

the task through multimodal QA, video-based QA, and graph-structured reasoning over perception, prediction, and planning. Recent benchmarks, including DriveLMM-o1 (Ishaq et al., 2025), NuPlanQA (Park et al., 2025), AutoDrive-QA (Khalili and Smyth, 2025), SURDS (Guo et al., 2025), STSBench (Fruhwirth-Reisinger et al., 2025), and STRIDE-QA (Ishihara et al., 2026), further emphasize multi-view perception and spatio-temporal reasoning. However, these benchmarks remain constrained to vehicle-mounted perspectives, limiting their ability to evaluate global intersection reasoning and infrastructure-side traffic understanding.

2.2 Roadside and Urban Outdoor VQA

Compared with ego-centric benchmarks, roadside VQA focuses on fixed-view scene understanding, enabling better modeling of global traffic dynamics and multi-agent interactions. SUTD-TrafficQA (Xu et al., 2021) formulates traffic event understanding as QA. TUMTraffic-VideoQA (Zhou et al., 2025) and RoadSceneVQA (Guan et al., 2026) extend this paradigm to roadside scenarios with video QA, grounding, and higher-level semantics. Meanwhile, City-3DQA (Sun et al., 2024) and Open3D-VQA (Zhang et al., 2025) extend QA to urban environments from a drone-based perspective. However, existing benchmarks still rely on single-modality observations and lack explicit 3D geometric information, limiting their ability to support multimodal reasoning that integrates multi-directional views and 3D spatial structure.

2.3 Vision-Language Models for Traffic Scene

Existing traffic VLMs for scene understanding span camera-only, LiDAR-only, and multimodal paradigms. Camera-based models such as OmniDrive (Wang et al., 2025) and OpenDriveVLA (Zhou et al., 2026) rely on image inputs, while TraffiX-Qwen (Zhou et al., 2025) extends them to roadside scenarios. LiDAR-based and multimodal methods such as LiDAR-LLM (Yang et al., 2025), NuScenes-QA (Qian et al., 2024), and BEV-LLM (Brandstätter et al., 2025) incorporate LiDAR inputs but are designed for ego-centric settings. In contrast, roadside models remain camera-based, limiting their ability to capture global 3D structure and spatial relationships. Our Inter-Geo addresses this gap by combining multi-view images with global point clouds, enabling cross-view semantics and geometry-aware reasoning.

3 Inter-3D VQA Dataset

3.1 Dataset Construction

As shown in Fig. 3, inspired by prior VQA dataset construction pipelines (Qian et al., 2024; Sima et al., 2024), we leverage open-source roadside perception data (Ding et al., 2026) to develop an automated QA generation framework that combines template filling with AI-assisted annotation and validation. All data are de-identified by anonymizing faces and license plates using automatic detection,

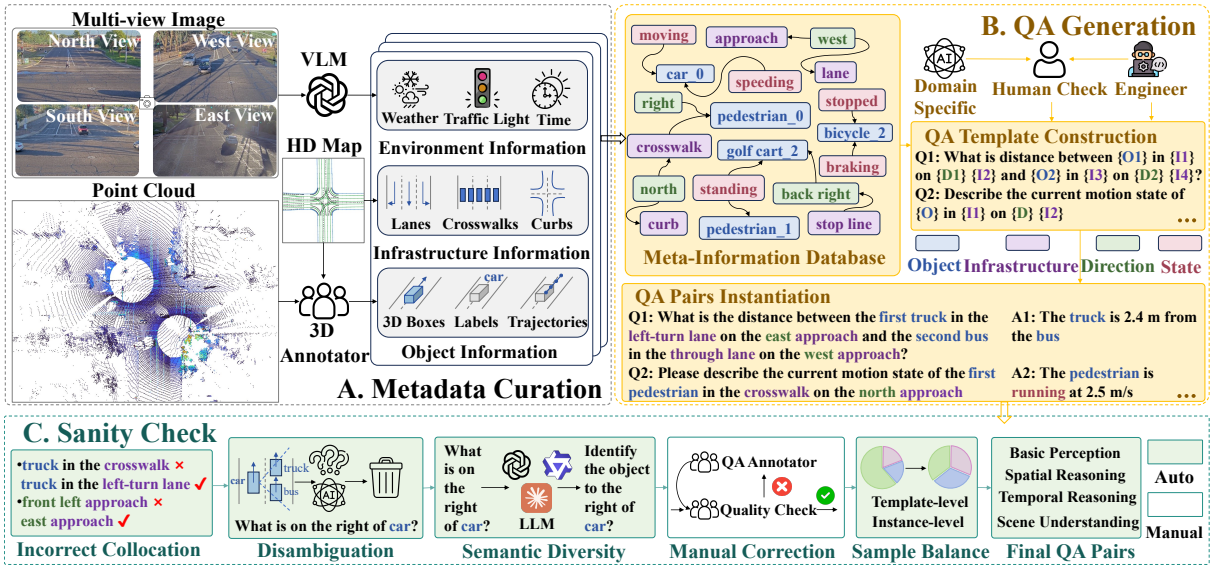


Figure 3: Data construction pipeline of Inter-3D VQA. The pipeline consists of three stages: metadata curation from synchronized multi-view images, LiDAR point clouds, and HD maps, template-based QA generation with LLM-assisted expansion, and a multi-stage sanity check for quality control and data balancing.

manual verification, and Gaussian blurring.

Metadata Curation. We utilize synchronized multi-view video and LiDAR point cloud streams, sampling keyframes at 2 Hz to construct image–point cloud pairs. After manual filtering, the raw data covers diverse environmental conditions, traffic patterns, and challenging perception scenarios. We then extract structured metadata from these multimodal inputs. Environmental semantics (e.g., traffic light state) are extracted from images using VLMs, while LiDAR data with annotated 3D bounding boxes provides object-level attributes and temporal associations. In addition, High-Definition maps are aligned with point clouds to extract infrastructure elements, enabling topology-aware reasoning. Finally, all modalities are unified into a structured per-frame representation that captures comprehensive information about the environment, objects, and infrastructure.

QA Generation. To improve dataset diversity, we adopt a template-based generation strategy that combines expert knowledge with Large Language Models (LLMs). We first collect representative traffic questions from domain experts and use a domain-adapted LLM to generate derivative questions. To ensure quality, the generated questions are manually curated and grouped into standardized templates with reserved metadata slots. Unlike conventional approaches, our templates encode precise 3D spatial and infrastructure topology information. Finally, QA pairs are instantiated by filling templates with structured metadata.

Sanity Check. We employ a multi-stage quality control pipeline to ensure the accuracy and diversity of QA pairs, with human intervention only at key stages. QA pairs first undergo consistency checks to remove instances violating common sense or traffic rules, followed by ambiguity filtering to discard questions with multiple valid answers or unclear semantics. LLMs then generate semantically equivalent variants to enhance linguistic diversity. A sampled manual review corrects low-quality samples, and finally template- and instance-level sampling are applied to balance the distribution across task categories.

3.2 Object Referring

Existing roadside VQA datasets typically associate objects with language through structured grounding or free-form descriptions, both aiming to achieve unambiguous object references. Grounded representations, as used in TUMTraffic-VideoQA (Zhou et al., 2025), support object localization and temporal correspondence. However, they rely on single-view observations from 2D images, which are insufficient to support spatial reasoning with accurate 3D information. Free-form descriptions, as used in RoadSceneVQA (Guan et al., 2026), offer greater flexibility but often introduce ambiguity when multiple similar objects are present, leading to degraded model performance.

To address these limitations, Inter-3D VQA ex-

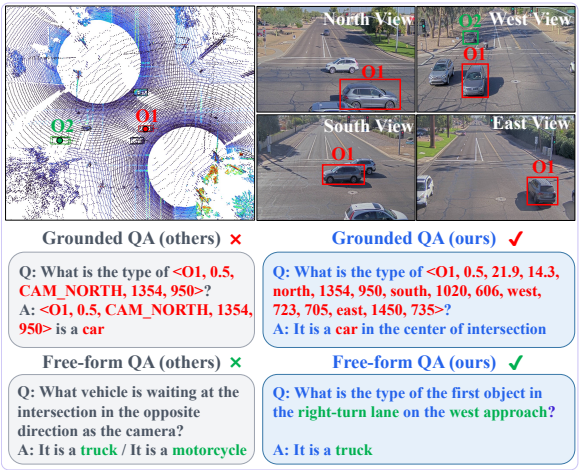


Figure 4: Object referring mechanisms in Inter-3D VQA. Grounded QA provides spatiotemporal object references for cross-view and 3D grounding, while free-form QA uses topology-aware language descriptions to reduce ambiguity through infrastructure constraints.

tends prior object referring schemes (Park et al., 2025) into two complementary QA formats as shown in Fig. 4. Free-form QA uses natural language descriptions enriched with infrastructure topology, enabling flexible reasoning over lanes, crosswalks, road structures, and traffic participants while reducing ambiguity among similar objects. Grounded QA provides explicit object references through a 3D-aware spatiotemporal tuple $< o, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, \ldots >$, where o denotes a unique object identifier, f_n is the temporal offset, (x_{3d}, y_{3d}) represents the 3D object center, and $(cam_{view}, x_{2d}, y_{2d}, \ldots)$ encodes multi-view projections. This representation aligns point clouds, multi-view images, language references, and infrastructure topology, enabling precise cross-view association, 3D localization, temporal tracking, and topology-aware object disambiguation.

3.3 Task Definition

Existing roadside VQA benchmarks (Zhou et al., 2025; Guan et al., 2026) focus on image-based tasks, making it difficult to support precise 3D measurements (e.g., distance to a crosswalk), infrastructure-aware relations (e.g., lane-level topology), and interaction reasoning (e.g., potential hazards). To address this gap, Inter-3D VQA integrates multi-view images with global point clouds and organizes the benchmark into four task categories: Basic Perception, Spatial Reasoning, Temporal Reasoning, and Scene Understanding. These tasks cover capabilities from low-level perception to high-level reasoning in complex roadside scenarios. Inter-3D VQA contains 407K QA pairs, split into training and test sets at an 8:2 ratio with no scene overlap to prevent memorization bias. Although certain safety-critical events can be rare in a single-intersection case, our QA construction pipeline explicitly covers diverse traffic behaviors and safety-relevant interactions. Queue dynamics, multi-agent conflicts, and abnormal events account for 16.8% of total QA pairs, while over 30% involve speeding or lane-changing behaviors rather than regular passing. Detailed task definitions and dataset statistics are provided in the appendix.

3.4 Inter-Metrics

Existing roadside VQA benchmarks also exhibit limitations in evaluating 3D-grounded reasoning. RoadSceneVQA (Guan et al., 2026) relies on language generation metrics, which fail to capture fine-grained semantic differences (e.g., opposite traffic directions may yield similar scores). TUMTraffic-VideoQA advances roadside video understanding through spatio-temporal video grounding, but its grounding remains primarily in the image/video space, making it difficult to assess whether a model understands real-world 3D positions, distances, trajectories, and safety-critical interactions. For 3D-grounded traffic VQA, correct language is not enough. A model must also be numerically accurate in physical space and semantically correct with respect to traffic structure. We therefore propose Inter-Metrics, a unified framework that evaluates models across textual consistency, numerical accuracy, and semantic correctness.

Textual Consistency. For open-ended textual answers, we measure similarity between the predicted answer $\hat{y}$ and the reference y using multiple metrics (BLEU-4 (2002), ROUGE-L (2004), BERTScore (2020), and SimCSE (2021)), aggregated as:

$$S_{\text{text}}(\hat{y}, y) = \sum_{m \in \mathcal{M}} w_m \, s_m(\hat{y}, y) \tag{1}$$

where s_m denotes the score of metric m, and w_m is its normalized weight satisfying $\sum_{m \in \mathcal{M}} w_m = 1$.

Numerical Accuracy. We evaluate scalar fields $\mathcal{S}$ (e.g., count, speed, distances) using absolute error and vector fields $\mathcal{V}$ (e.g., 3D coordinates, trajectories) using Euclidean error:

$$e_i = \begin{cases} |\hat{v}_i - v_i|, & \text{if } i \in \mathcal{S}, \\ \|\hat{\mathbf{p}}_i - \mathbf{p}_i\|_2, & \text{if } i \in \mathcal{V}, \end{cases} \tag{2}$$

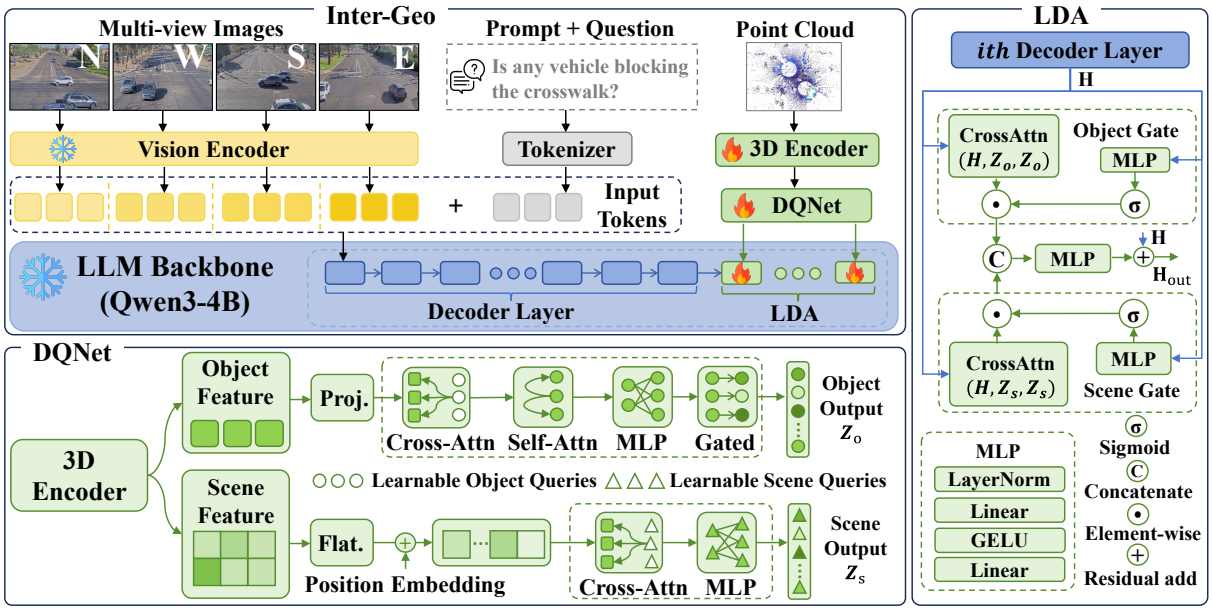


Figure 5: Architecture of Inter-Geo model, which integrates multi-view images, text, and global point clouds. Object- and scene-level features are extracted by 3D encoder, then transformed into dual aligned LiDAR representations via DQNet, and finally injected into the LLM decoder through LDA for geometry-aware reasoning.

To ensure comparability across heterogeneous quantities with different units and scales, errors are normalized by field-specific thresholds τ_i and mapped to a unified score in the range $[0, 1]$:

$$S_{\text{num}} = 1 - \frac{1}{N} \sum_{i=1}^{N} \min\left(\frac{e_i}{\tau_i}, 1\right) \tag{3}$$

where N is the number of evaluated fields. The same field-specific thresholds are applied consistently across all models to ensure a fair comparison.

Semantic Correctness. We evaluate semantic correctness by exact matching of structured semantic fields (Antol et al., 2015), such as object categories, lane membership, movement direction, traffic states, and conflict labels, going beyond surface-level linguistic similarity:

$$S_{\text{sem}}(\hat{\mathbf{z}}, \mathbf{z}) = \frac{1}{|\mathcal{F}|} \sum_{f \in \mathcal{F}} \mathbb{I}(\hat{z}_f = z_f) \tag{4}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and $\hat{z}_f$ and z_f denote the predicted and ground-truth values of semantic field f, respectively.

4 Methods

4.1 Overall Architecture

As shown in Fig. 5, we propose Inter-Geo, a LiDAR-enhanced vision-language model built on Qwen3-VL-4B-Instruct. Following the Flamingo design (Alayrac et al., 2022), we preserve the original camera–text pipeline and integrate LiDAR modality through decoder-level cross-attention, enabling geometry-aware reasoning without disrupting pretrained multimodal alignment. The model takes multi-view images $\mathcal{I} = \{I_N, I_W, I_S, I_E\}$ and text $\mathcal{T}$ as inputs. Images are encoded into visual embeddings $V = E_{\text{vis}}(\mathcal{I})$, which are inserted into the token sequence with text embeddings $T = E_{\text{tok}}(\text{Tokenizer}(\mathcal{T}))$, forming the multimodal input $H^0 = [T; V]$ for the decoder.

In parallel, LiDAR input is processed by 3D encoder to extract object- and scene-level features $\{X_o, X_s\}$, which the Dual Querying Net converts into aligned LiDAR representations:

$$(Z_o, Z_s) = \text{DQNet}(X_o, X_s) \tag{5}$$

These representations are injected into selected decoder layers through LiDAR Decoder Adapter, where hidden states are updated via cross-attention:

$$H^{l+1} = \text{LDA}^l(\text{Dec}^l(H^l), Z_o, Z_s) \tag{6}$$

4.2 3D Encoder

Our framework is agnostic to 3D perception backbone. In practice, we adopt a LION-based backbone (Liu et al., 2024) with a TransFusion detection head (Bai et al., 2022), though any 3D detection model can be used. Given point clouds $\mathcal{P}$, the

Types	Models	Free-form QA			Grounded QA		
		$S_{\text{text}} \uparrow$	$S_{\text{num}} \uparrow$	$S_{\text{sem}} \uparrow$	$S_{\text{text}} \uparrow$	$S_{\text{num}} \uparrow$	$S_{\text{sem}} \uparrow$
Open-source Generalist VLMs	LLaVA-NeXT-8B (2024)	0.788	0.347	0.646	0.803	0.341	0.656
	Llama3.2-11B (2024)	0.770	0.289	0.583	0.788	0.364	0.576
	Qwen3-VL-4B (2025)	0.806	0.397	0.709	0.827	0.432	0.736
	Qwen3-VL-8B (2025)	**0.821**	0.457	**0.749**	0.851	0.488	0.802
	InternVL3-4B (2025)	0.794	0.363	0.656	0.811	0.405	0.666
	InternVL3-8B (2025)	0.814	0.435	0.724	0.841	0.484	0.768
Specialized Driving VLMs	Senna-VLM (2024)	0.784	0.321	0.628	0.808	0.402	0.644
	NuScenes-QA (2024)	0.671	0.288	0.294	0.612	0.127	0.173
	OmniDrive (2025)	0.752	0.336	0.560	0.731	0.286	0.440
	OpenDriveVLA (2026)	0.777	0.421	0.624	0.774	0.337	0.534
	DriveLM (2024)	0.766	0.381	0.582	0.755	0.272	0.481
	BEV-LLM (2025)	0.772	0.376	0.613	0.761	0.301	0.507
Specialized Roadside VLMs	TraffiX-Qwen (2025)	0.753	0.322	0.564	0.743	0.264	0.495
	Inter-Geo (ours)	0.820	**0.482**	0.748	**0.859**	**0.589**	**0.828**

Table 2: Performance comparison of Inter-Geo with various VLMs on both free-form and grounded Inter-3D VQA.

encoder extracts scene-level BEV features X_s capturing global context, and object-level features X_o describing individual instances. The subsequent DQNet operates solely on X_s and X_o, enabling flexible integration with different backbones.

4.3 Dual Querying Net

The Dual Querying Net (DQNet) converts variable-length LiDAR features into fixed-size representations aligned with the language model latent space (Li et al., 2023). It consists of an object branch and a scene branch, capturing instance-level details and global context.

In the object branch, we adopt DETR-inspired query aggregation (Wang et al., 2021), where learnable queries Q_a attend to projected object tokens T_o and produce fixed object-aware representations refined by attention and MLP layers:

$$\hat{Q}_a = \text{MLP}(\text{Attn}_{\text{self}}(\text{Attn}_{\text{cross}}(Q_a, T_o))) \quad (7)$$

A confidence gate is then applied to suppress unreliable or redundant representations:

$$Z_o = \sigma(W_c \hat{Q}_a) \odot \hat{Q}_a \quad (8)$$

where $\sigma(\cdot)$ denotes the sigmoid function and W_c is a learnable projection.

In the scene branch, BEV features are projected into the language model space as tokens T_s and compressed using a Perceiver-style latent bottleneck (Jaegle et al., 2022). Specifically, learnable queries Q_s aggregate global context via cross-attention and are refined as:

$$Z_s = \text{MLP}(\text{Attn}_{\text{cross}}(Q_s, T_s)) \quad (9)$$

4.4 LiDAR Decoder Adapter

The LiDAR Decoder Adapter (LDA) integrates LiDAR representations into the Qwen3-VL decoder (Alayrac et al., 2022). Instead of appending LiDAR features as input tokens, LDA enables selective retrieval of geometric information while preserving the pretrained language–vision alignment. Given decoder states H, scene- and object-level information are retrieved via cross-attention:

$$\Delta_{\text{s/o}} = \text{Attn}_{\text{cross}}(H, Z_{s/o}) \quad (10)$$

To regulate LiDAR contributions, we employ a token- and channel-wise gating mechanism (Perez et al., 2018), which adaptively scales LiDAR features based on the language hidden states:

$$G_{s/o} = \sigma(\text{MLP}_{\text{s/o}}(\text{H})) \quad (11)$$

The gated features are then fused and injected into the decoder states via a residual module:

$$H_{\text{out}} = H + \text{MLP}_f([G_s \odot \Delta_{\text{s}};\ G_o \odot \Delta_{\text{o}}]) \quad (12)$$

where $[\cdot;\cdot]$ represents feature concatenation.

5 Experiments

5.1 Experimental Settings

For Inter-Geo, we adopt Qwen3-VL-4B-Instruct as the base model and perform LoRA-based fine-tuning using Llama-Factory (Zheng et al., 2024). We train the model using a learning rate of 1×10^{-4} and cosine scheduling. It is trained for 3 epochs with a per-GPU batch size of 1 and gradient

3D Enc.	DQNet		LDA	$S_{\text{text}} \uparrow$	$S_{\text{num}} \uparrow$	$S_{\text{sem}} \uparrow$
	Obj.	Scene				
✓	✓	✓	✓	0.859	0.589	0.828
	✓	✓	✓	0.841	0.477	0.765
✓		✓	✓	0.857	0.566	0.822
✓	✓		✓	0.857	0.553	0.817
✓	✓	✓		0.827	0.432	0.736

Table 3: Contributions of LiDAR-related modules. Removing both DQNet branches is equivalent to removing LDA, as neither provides LiDAR information.

Backbone	$S_{\text{text}} \uparrow$	$S_{\text{num}} \uparrow$	$S_{\text{sem}} \uparrow$
Qwen3-VL-2B-Instruct	0.853	0.558	0.812
Qwen3-VL-4B-Instruct	0.859	0.589	0.828
Qwen3-VL-8B-Instruct	0.869	0.617	0.848

Table 4: Effect of backbone model scale on Inter-Geo.

Injection	Stage	$S_{\text{text}} \uparrow$	$S_{\text{num}} \uparrow$	$S_{\text{sem}} \uparrow$
Input Embed	Input	0.856	0.552	0.816
Vision Fusion	Input	0.850	0.549	0.810
LDA	Decoder	0.859	0.589	0.828

Table 5: Effect of LiDAR modality injection strategy.

accumulation of 8. All experiments are conducted on four NVIDIA RTX PRO 6000 Blackwell GPUs. For a fair comparison, all open-source baselines are fine-tuned and evaluated using the same data split, prompts, grounded references, and evaluation protocol. Additional implementation details are provided in the appendix.

5.2 Quantitative Results

Table 2 shows Inter-Geo achieves the best grounded QA performance among existing VLMs while remaining highly competitive on free-form QA. For free-form QA, Inter-Geo obtains the highest numerical score $S_{\text{num}} = 0.482$, outperforming the strongest open-source baseline Qwen3-VL-8B by 5.5%. Its textual and semantic scores are nearly identical to Qwen3-VL-8B, indicating that Inter-Geo preserves strong language understanding while improving geometric reasoning through aligned LiDAR representations. For grounded QA, Inter-Geo outperforms all baselines across all metrics. Compared to Qwen3-VL-8B, it improves numerical accuracy by 20.7%. Compared with the top-performing driving-specific model, Senna-VLM, it improves S_{num} and S_{sem} by 46.5% and 28.6%. These gains suggest that the dual-branch LiDAR representations with decoder-level injection are effective for grounded QA that requires 3D localization and spatial reasoning. Although TraffiX-Qwen is roadside-specific, its image-dominant design relies on 2D cues for depth and distance reasoning, which can be unreliable under occlusion and cross-view ambiguity.

5.3 Ablation Studies

Effect of LiDAR-related Modules. Table 3 evaluates LiDAR-related modules on grounded Inter-3D VQA. Replacing the 3D encoder with a MLP encoder reduces S_{num} by 19.0% and S_{sem} by 7.6%, showing that explicit 3D perception is essential for providing structured geometric inputs to DQNet and forming reliable object- and scene-level LiDAR representations.

Dual querying branches play complementary roles in DQNet. Removing object branch reduces S_{num} by 4.0%, indicating its contribution to instance-level grounding through object positions and local spatial relations. Removing scene branch causes a larger drop of 9.5%, indicating that global BEV context is crucial for lane topology, intersection layout, and long-range spatial reasoning. This explains why the dual-branch design mainly improves numerical QA, where accurate 3D spatial evidence is required.

LDA is the key interface between LiDAR representations and language reasoning. Removing LDA blocks decoder-level access to DQNet output and degrades the model to a vision-language baseline incapable of effectively capturing LiDAR information, leading to a 26.7% drop in S_{num} and a 11.1% drop in S_{sem}. This indicates that encoding and aligning LiDAR features alone is insufficient. They must be explicitly injected into decoder layers to support final language reasoning.

Effect of Different Backbone Scales. We evaluate Inter-Geo with backbones of different scales on grounded Inter-3D VQA. As shown in Table 4, scaling the backbone from 4B to 8B improves S_{text}, S_{num}, and S_{sem} by 1.2%, 4.7%, and 2.4%, respectively. Notably, even the 2B variant outperforms existing VLM baselines, suggesting that aligned LiDAR representations provide complementary 3D geometric cues beyond standard VLM capabilities. These results show that Inter-Geo benefits from stronger backbones while retaining our advantage in multimodal LiDAR integration design.

Effect of LiDAR Modality Injection Strategy. In Table 5, we compare two alternative LiDAR

Model	$S_{\text{text}} \uparrow$	$S_{\text{num}} \uparrow$	$S_{\text{sem}} \uparrow$
Qwen3-4B	0.769 (-10.5%)	0.365 (-38.0%)	0.510 (-38.4%)
Qwen3-8B	0.783 (-8.8%)	0.411 (-30.2%)	0.573 (-30.8%)

Table 6: Performance of text-only LLMs on the grounded Inter-3D VQA, with relative performance drops compared to Inter-Geo.

Model	Inference Speed
Qwen3-4B	3.1 FPS
Qwen3-VL-4B	2.6 FPS
Inter-Geo	1.3 FPS

Table 7: Inference speed comparison of Inter-Geo with text-only LLM and image-only VLM baselines.

injection strategies with LDA. The input-embed strategy appends LiDAR embeddings to the input sequence as `<image_embeds>` + `<lidar_embeds>` + `<text_tokens>`. Although this improves over image-only VLMs, it treats LiDAR features as ordinary tokens, forcing geometric cues to compete with image and text tokens in the same attention space. The vision fusion strategy injects LiDAR into image embeddings through cross-attention, forming `<lidar-enhanced image_embeds>` + `<text_tokens>`. This early-fusion strategy may disrupt pretrained visual representations and dilute independent object- and scene-level geometric structures. In contrast, LDA allows the decoder to selectively retrieve 3D cues conditioned on language states, preserving pretrained image–text alignment while improving semantic grounding and spatial reasoning.

Effect of Visual Information. We additionally evaluate text-only LLMs on the grounded Inter-3D VQA to assess the contribution of visual information. As shown in Table 6, text-only models consistently underperform vision-language models and Inter-Geo, with particularly large performance gaps on numerical and semantic metrics. These results indicate that language priors alone are insufficient for grounded roadside reasoning and highlight the importance of incorporating visual and geometric information.

Runtime and Deployment Analysis. To assess inference speed, we compare Inter-Geo with text-only and image-only models using the same backbone as shown in Table 7. All models are evaluated on a single GPU with batch size 1. While Inter-Geo incurs additional computational cost, the text-only baseline reaches only 3.1 FPS, indicating that inference efficiency remains a broader challenge. In practice, sensing, preprocessing, and data transmission further contribute to end-to-end latency, motivating future work on more efficient deployment.

6 Conclusion

We introduce Inter-3D VQA, a roadside multimodal benchmark integrating multi-view images, global point clouds, and infrastructure topology for spatiotemporal reasoning. We further develop Inter-Geo, a LiDAR-enhanced MLLM that serves as a baseline for multimodal roadside reasoning. Experiments demonstrate the value of incorporating explicit LiDAR geometry, particularly for grounded spatial and numerical reasoning, while controlled ablations evaluate the contribution of the proposed fusion design. We hope Inter-3D VQA provides a foundation for future research on 3D-aware MLLMs for roadside traffic understanding.

Limitations

Although we introduce Inter-3D VQA, a large-scale roadside multimodal benchmark for 3D reasoning, and Inter-Geo, a multimodal large language model that incorporates LiDAR information with multi-view images, there are still several limitations.

First, while the results suggest that the 3D VQA tasks introduced in our benchmark are challenging for current models, the data underlying the benchmark is collected from a single intersection with a fixed geometry and layout. Cross-intersection validation is therefore necessary in future work. Second, Inter-Metrics depends on structured parsing of model outputs. When responses deviate from expected formats, semantic parsing may fail and affect S_{sem}, even if part of the answer is reasonable. Third, Inter-Geo mainly uses single-frame multimodal inputs and does not fully exploit multi-frame temporal cues, limiting its ability to model long-term motion patterns and dynamic interactions. Future work will expand scenario diversity, improve robust semantic parsing, and develop multi-frame

LiDAR-image reasoning models.

Ethical Considerations

As MLLMs are increasingly explored for ITS applications, greater attention should be given to how their outputs are interpreted and used in practice. Inaccurate outputs may lead to suboptimal policy or operational decisions, such as unnecessary signal-timing changes, or overly conservative roadside safety warnings. Model performance may also vary across regions with different traffic rules, roadway designs, and population characteristics. These considerations highlight the importance of context-specific validation and appropriate use of MLLMs as decision-support tools.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Anthropic. 2025. Claude Haiku 4.5. Accessed: Aug. 20, 2026.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. 2022. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 1080–1089.

Felix Brandstätter, Erik Schütz, Katharina Winter, and Fabian B. Flohr. 2025. BEV-LLM: leveraging multimodal BEV maps for scene captioning in autonomous driving. In *IEEE Intelligent Vehicles Symposium, IV 2025, Cluj-Napoca, Romania, June 22-25, 2025*, pages 345–350.

Haohan Chi, Huan-ang Gao, Ziming Liu, Jianing Liu, Chenyu Liu, Jinwei Li, Kaisen Yang, Yangcheng Yu, Zeda Wang, Wenyi Li, Leichen Wang, Xingtao Hu, Hao Sun, Hang Zhao, and Hao Zhao. 2025. Impromptu VLA: open weights and open data for driving vision-language-action models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.

Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie-Francine Moens. 2019. Talk2car: Taking control of your self-driving car. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2088–2098.

Shaozu Ding, Linan Song, Marco De Vincenzi, and Dajiang Suo. 2026. Resolve: A multi-resolution and multi-modal dataset for roadside cooperative perception. *Preprint*, arXiv:2606.31895.

Christian Fruhwirth-Reisinger, Dusan Malic, Wei Lin, David Schinagl, Samuel Schulter, and Horst Possegger. 2025. Stsbench: A spatio-temporal scenario benchmark for multi-modal large language models in autonomous driving. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6894–6910.

Google. 2025. Gemini 2.5 Flash. Accessed: Aug. 20, 2026.

Runwei Guan, Rongsheng Hu, Shangshu Chen, Ningyuan Xiao, Xue Xia, Jiayang Liu, Beibei Chen, Ziren Tang, Ningwei Ouyang, Shaofeng Liang, Yuxuan Fan, Wanjie Sun, and Yutao Yue. 2026. Roadscenevqa: Benchmarking visual question answering in roadside perception systems for intelligent transportation system. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 4366–4375.

Xianda Guo, Ruijun Zhang, Yiqun Duan, Yuhang He, Dujun Nie, Wenke Huang, Chenming Zhang, Shuai Liu, Hao Zhao, and Long Chen. 2025. SURDS: benchmarking spatial understanding and reasoning in driving scenarios with vision language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.

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 H. Khan. 2025. Drivelmm-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2025, Hangzhou, China, October 19-25, 2025*, pages 20501–20508.

Keishi Ishihara, Kento Sasaki, Tsubasa Takahashi, Daiki Shiono, and Yu Yamaguchi. 2026. STRIDE-QA: visual question answering dataset for spatiotemporal reasoning in urban driving scenes. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 5257–5266.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J. Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. 2022. Perceiver IO: A general architecture for structured inputs & outputs. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. 2024. Senna: Bridging large vision-language models and end-to-end autonomous driving. *Preprint*, arXiv:2410.22313.

Boshra Khalili and Andrew W. Smyth. 2025. Autodrive-qa: A multiple-choice benchmark for vision-language evaluation in urban autonomous driving. *Preprint*, arXiv:2503.15778.

Alex H. Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 12697–12705.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: Stronger llms supercharge multimodal capabilities in the wild. Accessed: Aug. 20, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhe Liu, Jinghua Hou, Xinyu Wang, Xiaoqing Ye, Jingdong Wang, Hengshuang Zhao, and Xiang Bai. 2024. LION: linear group RNN for 3d object detection in point clouds. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Ana-Maria Marcu, Long Chen, Jan Hünermann, Alice Karnsund, Benoît Hanotte, Prajwal Chidananda, Saurabh Nair, Vijay Badrinarayanan, Alex Kendall, Jamie Shotton, Elahe Arani, and Oleg Sinavski. 2024. Lingoqa: Visual question answering for autonomous driving. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXVII*, volume 15135 of *Lecture Notes in Computer Science*, pages 252–269.

Meta. 2024. Llama 3.2 model card. Accessed: Aug. 20, 2026.

OpenAI. 2025. GPT-5.1 model. Accessed: Aug. 20, 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318.

Sung-Yeon Park, Can Cui, Yunsheng Ma, Ahmadreza Moradipari, Rohit Gupta, Kyungtae Han, and Ziran Wang. 2025. Nuplanqa: A large-scale dataset and benchmark for multi-view driving scene understanding in multi-modal large language models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pages 8066–8076.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and*

the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pages 3942–3951.

Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. 2024. Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 4542–4550.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. 2024. Drivelm: Driving with graph visual question answering. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LII*, volume 15110 of *Lecture Notes in Computer Science*, pages 256–274.

Penglei Sun, Yaoxian Song, Xiang Liu, Xiaofei Yang, Qiang Wang, Tiefeng Li, Yang Yang, and Xiaowen Chu. 2024. 3d question answering for city scene understanding. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 2156–2165.

Haiyang Wang, Chen Shi, Shaoshuai Shi, Meng Lei, Sen Wang, Di He, Bernt Schiele, and Liwei Wang. 2023. DSVT: dynamic sparse voxel transformer with rotated sets. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 13520–13529.

Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and José M. Álvarez. 2025. Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 22442–22452.

Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. 2021. DETR3D: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning, 8-11 November 2021, London, UK*, volume 164 of *Proceedings of Machine Learning Research*, pages 180–191.

Li Xu, He Huang, and Jun Liu. 2021. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9878–9888.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K. Wong, Zhenguo Li, and Hengshuang Zhao. 2024. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics Autom. Lett.*, 9(10):8186–8193.

Senqiao Yang, Jiaming Liu, Renrui Zhang, Mingjie Pan, Ziyu Guo, Xiaoqi Li, Zehui Chen, Peng Gao, Hongsheng Li, Yandong Guo, and Shanghang Zhang. 2025. Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. In *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pages 9247–9255.

Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. 2021. Center-based 3d object detection and tracking. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11784–11793.

Junwei You, Pei Li, Zhuoyu Jiang, Weizhe Tang, Zilin Huang, Rui Gan, Jiaxi Liu, Yan Zhao, Sikai Chen, and Bin Ran. 2026. V2x-qa: A comprehensive reasoning dataset and benchmark for multimodal large language models in autonomous driving across ego, infrastructure, and cooperative views. *Preprint*, arXiv:2604.02710.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6281–6290.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 11941–11952.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Weichen Zhang, Zile Zhou, Xin Zeng, Xuchen Liu, Jianjie Fang, Chen Gao, Jinqiang Cui, Yong Li, Xinlei Chen, and Xiao-Ping Zhang. 2025. Open3d-vqa: A benchmark for embodied spatial concept reasoning with multimodal large language model in open space. In *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27-31, 2025*, pages 12784–12791.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3:*

System Demonstrations), ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 400–410.

Xingcheng Zhou, Xuyuan Han, Feng Yang, Yunpu Ma, Volker Tresp, and Alois Knoll. 2026. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 13782–13790.

Xingcheng Zhou, Konstantinos Larintzakis, Hao Guo, Walter Zimmer, Mingyu Liu, Hu Cao, Jiajie Zhang, Venkatnarayanan Lakshminarasimhan, Leah Strand, and Alois Knoll. 2025. TUMTraf videoQA: Dataset and benchmark for unified spatio-temporal video understanding in traffic scenes. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

Appendix

A Inter-3D VQA Dataset

A.1 Task Definition

While the main paper introduces the overall task formulation, we provide additional details and representative examples here to better illustrate the scope and challenges of each task category.

Existing roadside VQA benchmarks (Zhou et al., 2025; Guan et al., 2026) are limited to image-based semantic understanding, which restricts their ability to support fine-grained spatial and structural reasoning. For example, questions such as "How far is a pedestrian from the nearest crosswalk?" or "What is the distance between a vehicle and the stop line?" require precise geometric measurements that are difficult to estimate reliably from 2D images alone. Similarly, queries like "How many queued vehicles are in the northbound left-turn lane?" or "Is there a pedestrian at the southeast curb ramp?" involve topology-aware reasoning, requiring both spatial localization and an understanding of infrastructure semantics. In addition, interaction-based questions, such as identifying potential conflicts or unsafe behaviors among traffic participants, require modeling dynamic relationships over time. To address these challenges, Inter-3D VQA extends image-based reasoning to multimodal reasoning over synchronized multi-view images and global point clouds. This enables models to perform unified reasoning over 3D geometry, infrastructure topology, and temporal dynamics within a consistent spatial coordinate system.

We categorize the benchmark into four complementary task types:

- Basic Perception: This category focuses on a model's ability to identify fundamental information, including object types, counts, visual attributes, and environmental conditions. While image-based models typically perform well on such tasks, incorporating point cloud data provides more spatial observations especially under occlusion or limited viewpoints.

- Spatial Reasoning: This category emphasizes the evaluation of a model's geometric understanding within a three-dimensional space. The model is required to estimate absolute distances between objects using both images and point clouds, while also understanding the topological relationships between objects and infrastructure elements (e.g., lanes and crosswalks) for more precise spatial reasoning.

- Temporal Reasoning: Using sequential multi-frame images and point clouds, this category evaluates a model's ability to capture dynamic traffic scenarios, including motion analysis and trajectory inference for individual objects, as well as interaction modeling among multiple objects (e.g., following distance estimation and conflict detection).

- Scene Understanding: From a global perspective, this category integrates multi-view images and global point clouds to evaluate a model's understanding of overall intersection traffic states, including congestion analysis, risk detection, and traffic management.

A.2 Dataset Statistics

As shown in Fig. 6, we analyze Inter-3D VQA from multiple perspectives, including question length distribution, task composition, and vocabulary statistics. The histogram of question lengths shows that most questions contain approximately 14–22 words, indicating a balance between concise formulation and sufficient semantic detail for complex reasoning. The task distribution demonstrates a relatively balanced composition across the four task categories, reducing bias toward specific reasoning patterns. In particular, Spatial Reasoning and Temporal Reasoning occupy a substantial proportion of the benchmark, highlighting the emphasis on geometry-aware and dynamic traffic understanding.

The vocabulary statistics further reveal rich semantic coverage related to traffic participants, infrastructure topology, motion states, and interaction patterns. High-frequency terms such as "lane", "crosswalk", "turn", and directional descriptors indicate that the dataset explicitly models topology-aware and infrastructure-centric reasoning.

We further examine the distribution of answer lengths and their semantic content, assess object coverage, and analyze the distribution of numerical attributes. As shown in Fig. 7a, the distribution of answer lengths reveals that most responses remain concise, indicating that the benchmark prioritizes precise argumentation and reasoning rather than the generation of verbose answers. The word cloud in Fig. 7b highlights high-frequency terms related to traffic participants, road structures, motion states,

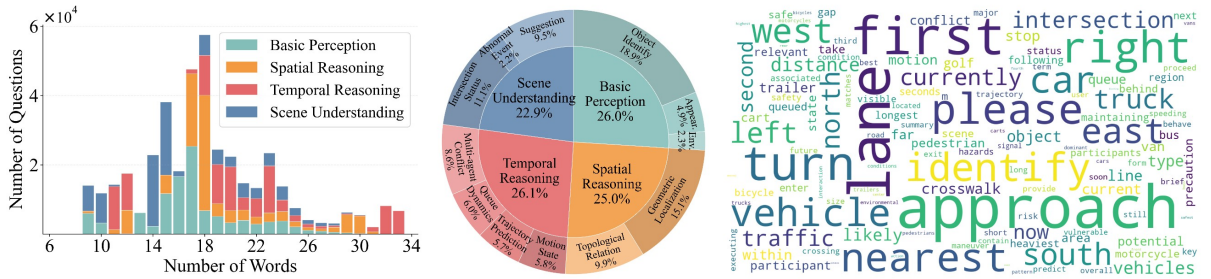
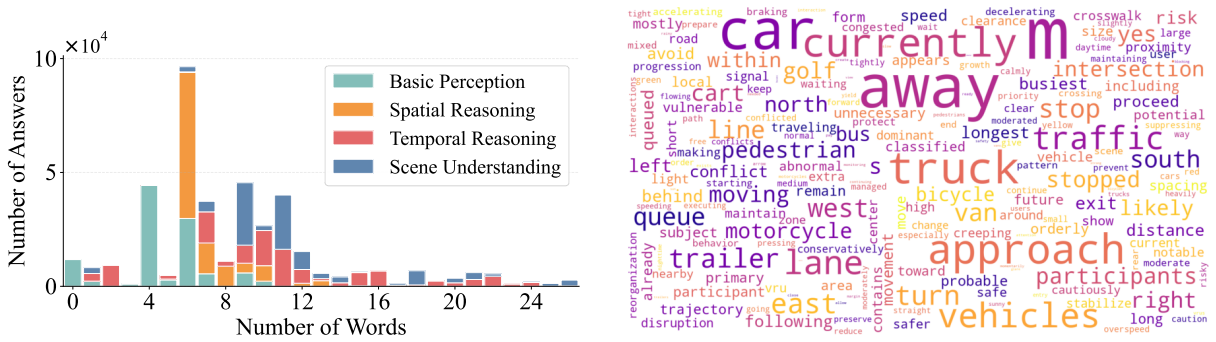


Figure 6: Question-level statistics of Inter-3D VQA. The dataset exhibits a balanced distribution across task categories and provides rich topology-aware descriptions for 3D reasoning.



(a) (b)

Figure 7: Answer-level statistics of Inter-3D VQA. (a) Distribution of answer lengths, demonstrating concise and reasoning-oriented responses. (b) Word cloud of answers, revealing dominant semantic concepts in traffic participants, infrastructure, and motion.

and decision-making cues, underscoring the close interplay between perception and reasoning within intersection scenarios.

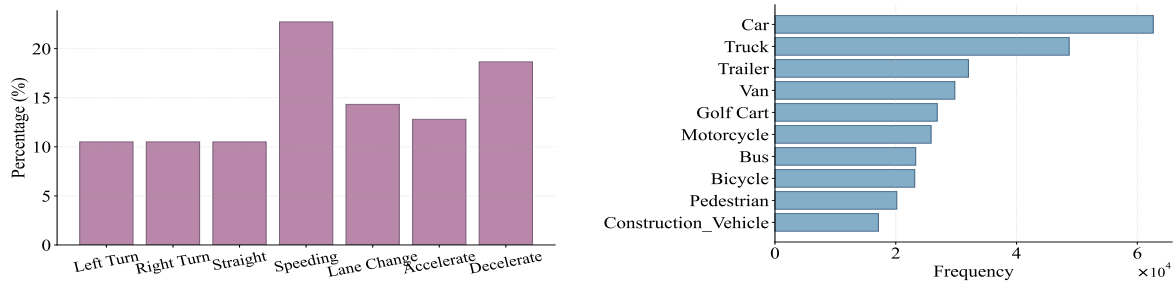
Fig. 8a shows a diverse distribution of driving behaviors, including turning, lane changes, acceleration, and speeding, enabling evaluation across both routine and safety-critical situations. Meanwhile, Fig. 8b demonstrates broad object coverage, including common vehicles as well as vulnerable and less frequent traffic participants, supporting diverse and realistic evaluation settings.

We also analyze the distributions of numerical targets and 3D spatial coordinates in QA pairs. As shown in Fig. 9, statistics of counts, distances, and velocities indicate that the dataset provides rich quantitative supervision, supporting tasks such as quantity estimation, geometric reasoning, and motion analysis. Fig. 10 shows that the distribution of 3D coordinates involved in QA pairs spans a wide range of intersection areas, indicating a wide spatial coverage rather than concentration in limited regions.

A.3 Inter-Metrics Details

We provide additional implementation details and evaluation procedures for Inter-Metrics. The framework is designed to jointly evaluate linguistic quality, geometric reasoning, and semantic correctness in Inter-3D VQA tasks.

Textual Consistency. We evaluate the generated answers using a weighted combination of lexical and semantic similarity metrics: Normalized Exact Match (NEM), BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2020), and SimCSE (Gao et al., 2021) similarity. For reproducibility, we specify the implementation details and parameter settings for each metric. BLEU-4 is computed using `sacrebleu==2.4.0` with the 13a tokenizer, case-sensitive scoring, exponential smoothing, and corpus-level aggregation. ROUGE-L is computed using `rouge-score==0.1.2`, and we report ROUGE-L F1 with `use_stemmer=True`. BERTScore is computed using `bert-score==0.3.13`, and we report the F1 score with roberta-large, `num_layers=17`, `idf=False`, and `rescale_with_baseline=True`. SimCSE similarity is computed using the



(a) (b)

Figure 8: Behavior and object statistics of Inter-3D VQA. (a) Distribution of driving behaviors, covering both routine and safety-critical actions. (b) Distribution of object categories, including diverse traffic participants.

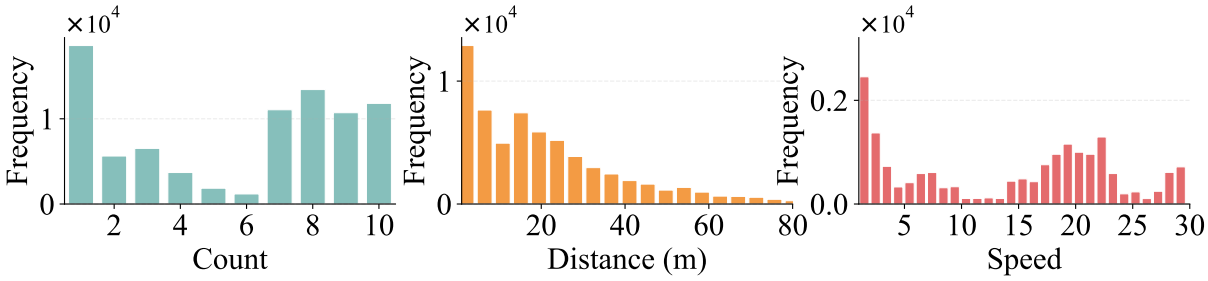


Figure 9: Distribution of numerical attributes in Inter-3D VQA, including object count, distance, and speed.

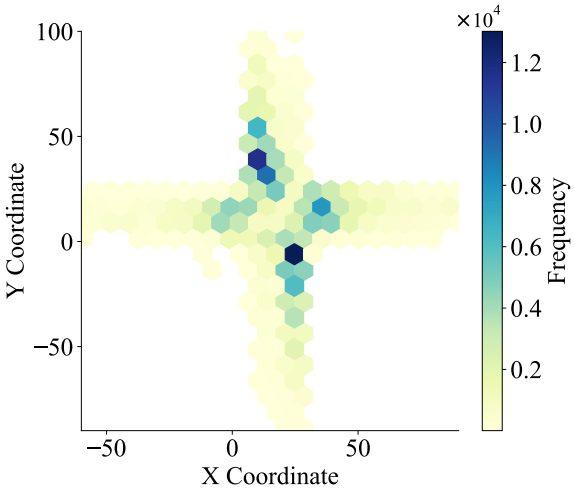


Figure 10: Spatial distribution of 3D coordinates in Inter-3D VQA.

`princeton-nlp/sup-simcse-roberta-large` checkpoint, where predictions and references are embedded separately and compared using cosine similarity. The final score is averaged over all evaluation examples.

Compared with conventional language-based evaluation, Inter-Metrics places greater emphasis on semantic similarity to reduce sensitivity to wording variations. For example, "The vehicle is turning left" and "The car is making a left turn" may receive relatively low scores under lexical-overlap metrics such as BLEU or ROUGE, despite describing the same traffic behavior. By incorporating semantic similarity metrics such as BERTScore and SimCSE, Inter-Metrics better captures semantic equivalence.

Numerical Accuracy. Unlike image-based spatial evaluation used in previous roadside benchmarks, Inter-Metrics evaluates quantities directly in physically meaningful 3D space. Scalar quantities (e.g., object count, distance, speed, and acceleration) are evaluated using absolute error, while vector quantities (e.g., 3D coordinates, image-plane coordinates, and trajectory waypoints) are evaluated using Euclidean distance error. To ensure comparability across heterogeneous quantities, each error is normalized using a predefined threshold specific to the corresponding field. Thresholds are set to 2 for count, 10 m for distance, 5 m/s for speed, 1 m/s^2 for acceleration, 8 m for global 3D coordinates, 150 px for image-plane coordinates, and 5 m for waypoint offsets. This allows the framework to more reliably measure geometric understanding, spatial localization, and trajectory reasoning. Although metric calculation relies on structured output parsing, all models are supervised to follow a unified answer format, resulting in a high parsing success rate.

Semantic Correctness. Inter-Metrics additionally evaluates structured semantic fields extracted from model predictions, including object categories, lane functions, signal states, motion directions, interaction patterns, and risk-related attributes. Each field is evaluated through exact matching against the ground-truth structured annotations. This design explicitly distinguishes semantic correctness from surface-level linguistic similarity, preventing semantically incorrect yet linguistically plausible predictions from receiving overly high scores. For example, "A vehicle on the northbound approach is turning left" and "A vehicle on the southbound approach is turning left" may obtain high lexical-overlap scores despite describing different traffic situations. Inter-Metrics explicitly evaluates structured semantic fields such as direction and interaction state, preventing semantically inconsistent predictions from being overestimated.

B Implementation Details

All models are evaluated on the same Inter-Metrics protocol. We use the same train/validation split across methods, with scene-level splitting and identical evaluation scripts. All reported results are averaged over five independent runs. Input camera order is fixed to `north`, `south`, `east`, and `west`.

B.1 Inter-Geo

We provide the main architectural settings of Inter-Geo. The model uses two LiDAR input branches: object-level features and scene-level LiDAR features. Raw object and scene features have dimensions 523 and 386, respectively, and are projected into a shared LiDAR hidden space of dimension 512. For numeric object attributes, we additionally employ label embeddings with vocabulary size 16 and embedding dimension 32.

To regulate LiDAR information injection, Inter-Geo retains at most 16 object representations, while scene-level inputs are truncated to 256 tokens and compressed into 64 scene representations before interacting with the decoder. Cross-modal attention in the LiDAR branch uses 8 attention heads. The top 4 decoder layers are further equipped with LiDAR Decoder Adapters (LDA), each using 8 attention heads. To stabilize early training, the adapter gates are initialized with bias -5.0, suppressing excessive LiDAR influence at initialization.

For optimization, Inter-Geo adopts LoRA fine-tuning on the Qwen3-VL-4B backbone with rank 8, scaling factor 16, and dropout 0.05. The image resolution is capped at 262,144 pixels, and the maximum sequence length is set to 4096.

B.2 Open-source Generalist VLMs

All open-source VLMs are trained with BF16 precision using a cosine learning rate schedule and a base learning rate of 1×10^{-4}, taking four-view images as input. Qwen3-VL (Bai et al., 2025) and InternVL (Zhu et al., 2025) models use a maximum image resolution of 262,144 pixels and sequence length of 2,048. The 4B variants adopt standard LoRA with rank 8, while the 8B variants employ rsLoRA with rank 32, $\alpha = 64$, and dropout 0.05. These models are trained with batch size 2 and gradient accumulation 4.

LLaVA-NeXT-8B (Li et al., 2024) uses a lower image resolution cap of 196,608 pixels and a longer sequence length of 8,192, while Llama-3.2-11B-Vision-Instruct (Meta, 2024) uses a resolution of 262,144 pixels and sequence length of 2,048. Both models are trained with batch size 1 and gradient accumulation 8.

B.3 Specialized Driving VLMs

We follow the original training and inference protocols of all specialized driving VLMs, adapting only the data interfaces for Inter-3D VQA. DriveLM (Sima et al., 2024) retains its LLaMA-Adapter-based 7B multimodal backbone and is trained for 3 epochs with batch size 4, effective learning rate 6.25×10^{-5}, and weight decay 0.05. Senna (Jiang et al., 2024) adopts LoRA fine-tuning for 3 epochs using batch size 1, gradient accumulation 8, and learning rate 1×10^{-4}. NuScenes-QA (Qian et al., 2024) preserves its MCAN-style (Yu et al., 2019) structured reasoning architecture and object-centric pipeline, trained for 4 epochs with batch size 32 and learning rate 3×10^{-4}.

OmniDrive (Wang et al., 2025) maintains its camera-only multi-view setting with four fixed views and is trained for 3 epochs using global batch size 12 and learning rate 2×10^{-5}. OpenDriveVLA (Zhou et al., 2026) preserves its original feature extraction and language generation pipeline with fixed camera ordering, trained using batch size 1, gradient accumulation 8, and learning rate 1×10^{-4}. BEV-LLM retains its "BEV Features $\rightarrow$ Q-Former $\rightarrow$ LLM" architecture and applies LoRA fine-tuning with rank 16, batch size 2, and learning rate 1×10^{-4} for 3 epochs.

Models	Free-form QA			Grounded QA		
	$S_{\text{text}} \uparrow$	$S_{\text{num}} \uparrow$	$S_{\text{sem}} \uparrow$	$S_{\text{text}} \uparrow$	$S_{\text{num}} \uparrow$	$S_{\text{sem}} \uparrow$
GPT-5.1 (2025)	0.715	0.251	0.407	0.523	0.081	0.098
Gemini-2.5-flash (2025)	0.725	0.204	0.428	0.534	0.141	0.125
Claude-4.5-Haiku (2025)	0.669	0.197	0.335	0.537	0.075	0.077

Table 8: Zero-shot performance of closed-source API-only VLMs on free-form and grounded Inter-3D VQA.

Model	Backbone	$S_{\text{text}} \uparrow$	$S_{\text{num}} \uparrow$	$S_{\text{sem}} \uparrow$
CenterPoint (Yin et al., 2021)	Sparse Convolution	0.862	0.540	0.828
TransFusion-L (Bai et al., 2022)	Sparse Convolution	0.857	0.580	0.819
DSVT (Wang et al., 2023)	Transformer	0.855	0.544	0.812
LION (Liu et al., 2024)	Mamba	0.859	0.589	0.828

Table 9: Comparison of different 3D encoders.

B.4 Specialized Roadside VLMs

TraffiX-Qwen (Zhou et al., 2025) adopts a 3-frame $\times$ 4-view configuration, yielding 12 images per QA sample. It uses Qwen2-0.5B with a SigLIP (Zhai et al., 2023) vision encoder as the backbone and is trained for 3 epochs using LoRA-based adaptation, with batch size 1, gradient accumulation 4, and learning rate 5×10^{-6}.

B.5 Closed-source API-only VLMs

Closed-source models are evaluated in a pure zero-shot setting without fine-tuning, using the same input samples and evaluation scripts as open-source models. We evaluate GPT-5.1 (OpenAI, 2025), Gemini-2.5-Flash (Google, 2025), and Claude Haiku 4.5 (Anthropic, 2025), all of which take four input images per sample. GPT uses low-detail image mode with reasoning disabled and a maximum of 512 output tokens. Gemini adopts temperature 0, `MEDIA_RESOLUTION_LOW`, and 512 maximum output tokens, with `thinkingBudget` set to 0 for Gemini-2.5-Flash. Claude uses temperature 0, resizes images to a maximum long edge of 1280 pixels, and limits outputs to 128 tokens.

C Benchmark Results

C.1 Evaluation of API-only VLMs

We report the zero-shot performance of closed-source API-only VLMs using four-view image inputs in Table 8. Since these models are evaluated without supervised fine-tuning, they are not adapted to the answer formats and structured fields required by our benchmark. As a result, although they show moderate language understanding on free-form QA, their performance drops substantially on grounded QA. For free-form QA, Gemini-2.5-Flash (Google, 2025) achieves the best textual and semantic scores, while GPT-5.1 (OpenAI, 2025) obtains the highest numerical score. However, on grounded QA, the best numerical and semantic scores remain only 0.141 and 0.125, respectively.

This degradation is mainly caused by two factors. First, without task-specific training, API-only models struggle to follow the required structured response format, especially for grounded answers involving object references, coordinates, numerical values, and semantic fields. Second, their outputs are often more free-form and descriptive, which increases parsing failures in numerical and semantic evaluation. Therefore, even when a response contains partially relevant information, it may not be successfully parsed into the expected fields, leading to low S_{num} and S_{sem}.

C.2 Different 3D Encoders

We compare different 3D encoders within the Inter-Geo framework in Table 9. All variants outperform image-based VLM baselines, demonstrating that the proposed DQNet and LDA design is compatible with diverse LiDAR perception backbones while consistently preserving the advantage of explicit 3D modeling.

Among these encoders, the Mamba-based LION (Liu et al., 2024) achieves the best overall performance, indicating its strength in modeling long-range spatial dependencies in point

Insertion Position	$S_{\text{text}} \uparrow$	$S_{\text{num}} \uparrow$	$S_{\text{sem}} \uparrow$
First 4 layers	0.858	0.554	0.816
Middle 4 layers	0.858	0.552	0.822
Last 4 layers	0.859	0.589	0.828

Table 10: Effect of LDA insertion position in the decoder on Inter-Geo.

# LDA Layers	$S_{\text{text}} \uparrow$	$S_{\text{num}} \uparrow$	$S_{\text{sem}} \uparrow$
2	0.858	0.544	0.821
4	0.859	0.589	0.828
8	0.857	0.553	0.816

Table 11: Effect of the number of LDA layers in the decoder on Inter-Geo.

# Object Queries	$S_{\text{text}} \uparrow$	$S_{\text{num}} \uparrow$	$S_{\text{sem}} \uparrow$
8	0.861	0.569	0.827
16	0.859	0.589	0.828
24	0.858	0.545	0.818

Table 12: Effect of the number of object queries in DQNet.

clouds. This capability is particularly beneficial for global scene understanding and spatial reasoning. TransFusion-L (Bai et al., 2022) obtains a strong numerical score ($S_{\text{num}} = 0.580$), suggesting that its detection-oriented strategy and object-level representations are effective for distance estimation and localization, although this does not necessarily translate into the best structured semantic accuracy.

CenterPoint (Yin et al., 2021) achieves the highest textual score ($S_{\text{text}} = 0.862$) but lower numerical performance ($S_{\text{num}} = 0.540$), likely because sparse convolution is effective for object detection but less expressive for global geometric reasoning. DSVT (Wang et al., 2023) remains competitive, yet its numerical and semantic scores are below LION, suggesting that Transformer-based voxel modeling is slightly less effective for global point cloud patterns in roadside scenarios.

C.3 LDA Insertion Position

We study the effect of LDA insertion position in the 36-layer Qwen3-VL-4B-Instruct decoder in Table 10. Compared with inserting LDA into the last four layers, inserting it into the first four layers reduces S_{num} by 6.3% and S_{sem} by 1.4%. This is likely because early decoder layers mainly capture lexical patterns and low-level multimodal alignment. Injecting LiDAR representations too early may disturb the pretrained image–text representations and force geometric information to pass through all subsequent transformations, thereby weakening structured 3D cues.

Inserting LDA into the middle layers slightly improves semantic correctness compared with early insertion, but still yields a 6.3% lower numerical score than late insertion. At this stage, the decoder has begun to form higher-level representations, but its hidden states may not yet be sufficiently task-specific to retrieve the most relevant object- and scene-level LiDAR representations.

In contrast, the last four decoder layers are closer to answer generation and contain more task-conditioned semantic states. Injecting LiDAR representations at this stage allows the model to access geometric cues after the visual-language context has been formed, reducing interference with pretrained representations while directly supporting final reasoning.

C.4 Number of LDA Layers

We further investigate the effect of applying LDA to different numbers of decoder layers. As shown in Table 11, the 4-layer configuration achieves the best overall performance. Compared with the 2-layer setting, it improves S_{num} by 8.3% and S_{sem} by 0.9%, suggesting that an insufficient number of layers limits the model's ability to acquire 3D cues, whereas an appropriate number of LDA layers provides sufficient capacity to effectively capture LiDAR representations and integrate geometric cues into the language reasoning process.

However, increasing the number of LDA layers to 8 does not bring further gains. Instead, S_{num} decreases by 6.1% and S_{sem} by 1.4%. This indicates that excessive LiDAR injection may disturb pretrained language–vision alignment and introduce redundant geometric cues across decoder layers, causing the model to over-rely on LiDAR features.

C.5 Number of Object Queries

In Table 12, we evaluate the effect of varying the number of object queries in the object branch of DQNet. Reducing the number of queries to 8 mainly affects numerical reasoning, with S_{num} decreasing by 3.4%, suggesting that too few queries limit the model's capacity to capture instance-level geometric cues.

Models	Count ↓	Distance ↓	Speed ↓	Acceleration ↓	Coord_{3d} ↓	Coord_{2d} ↓	Waypoint ↓
GPT-5.1	0.95	0.91	1.00	1.00	0.78	0.93	0.86
Gemini-2.5-flash	0.97	0.91	0.51	0.99	0.85	0.94	0.84
Claude-4.5-Haiku	0.95	0.94	0.71	1.00	0.99	1.00	0.90
LLaVA-NeXT-8B	0.35	0.24	0.69	0.94	0.81	0.87	0.71
Llama3.2-11B	0.37	0.21	0.63	0.84	0.81	0.86	0.74
Qwen3-VL-4B	0.26	0.18	0.53	0.64	0.77	0.81	0.78
Qwen3-VL-8B	0.22	0.15	0.37	0.73	0.73	**0.73**	0.66
InternVL3-4B	0.28	0.20	0.54	0.74	0.80	0.89	0.72
InternVL3-8B	0.25	0.18	0.38	**0.54**	0.77	0.78	0.71
Senna-VLM	0.28	0.24	0.61	0.63	0.81	0.87	0.74
NuScenes-QA	0.89	0.83	0.67	1.00	1.00	0.97	0.76
OmniDrive	0.32	0.69	0.50	0.91	0.94	0.90	0.75
OpenDriveVLA	0.33	0.67	0.44	0.74	0.75	0.90	0.82
DriveLM	0.34	0.72	0.45	0.97	0.95	0.94	0.73
BEV-LLM	0.34	0.63	0.31	0.93	0.97	0.99	0.72
TraffiX-Qwen	0.37	0.69	0.49	1.00	0.91	0.89	0.80
Inter-Geo (ours)	**0.22**	**0.15**	**0.29**	0.59	**0.48**	0.76	**0.31**

Table 13: Detailed numerical error comparison on grounded Inter-3D VQA across different numerical categories.

# Scene Queries	S_{text} ↑	S_{num} ↑	S_{sem} ↑
32	0.859	0.567	0.824
64	0.859	0.589	0.828
128	0.858	0.545	0.820

Table 14: Effect of the number of scene queries in DQNet.

Increasing the number of queries to 24 further degrades performance, reducing S_{num} by 7.5% and S_{sem} by 1.2%. In complex roadside scenes, excessive queries may attend to duplicated or low-relevance objects, introducing redundant or noisy representations and weakening the compactness of geometric evidence. Therefore, 16 object queries provide the best balance between object coverage and representation quality.

C.6 Number of Scene Queries

We further analyze the effect of the number of scene queries in the scene branch of DQNet. As shown in Table 14, the default setting with 64 queries achieves the best balance between global context coverage and representation compactness.

Reducing the number to 32 decreases S_{num} by 3.7% and S_{sem} by 0.5%, suggesting that too few scene queries provide insufficient capacity to aggregate global BEV context, such as lane topology, intersection layout, and long-range spatial relationships.

Increasing the number to 128 causes a larger degradation, with S_{num} dropping by 7.5% and S_{sem} by 1.0%. This indicates that excessive scene queries may weaken the bottleneck effect of the scene branch and introduce redundant or noisy spatial cues, causing the decoder to attend to less relevant regions and reducing the effectiveness of metric reasoning and semantic grounding.

C.7 Numerical Error Analysis

Table 13 reports detailed numerical errors on grounded Inter-3D VQA across different numerical categories. Compared with image-based open-source, driving-specific, and roadside-specific VLMs, Inter-Geo achieves the best performance on most geometry-related metrics. In particular, it obtains the lowest errors on distance, 3D coordinates, and waypoint prediction, with errors of 0.15, 0.48, and 0.31, respectively. These results demonstrate the advantage of incorporating object- and scene-level aligned LiDAR representations, which provide explicit geometric cues for spatial estimation, 3D localization, and trajectory-related reasoning. This is especially beneficial for tasks where vision-only models often suffer from depth ambiguity and limited metric awareness.

However, Inter-Geo does not outperform all baselines across every category. For 2D coordinate errors, it remains competitive but is not clearly superior, which is expected because this metric depends more on image-plane localization than 3D spatial grounding. In addition, its acceleration error remains relatively high compared with several

System Prompt:

You are an AI assistant for traffic-scene analysis at a four-way intersection.

Inputs are synchronized at the same timestamp:
- four camera images: north_image, south_image, east_image, west_image
- LiDAR-derived object and scene evidence.

Evidence usage:
- north/south/east/west_image show the corresponding intersection approaches.
- Use both images and LiDAR evidence conservatively.
- Prefer LiDAR-derived evidence for 3D geometry, global position, spatial layout, occlusion, and distance reasoning; use images as auxiliary evidence.
- Prefer images for visible appearance, signal states, and cross-view visual context; use LiDAR as auxiliary evidence.
- Do not hallucinate unsupported objects, states, interactions, hidden signals, or invisible objects.

Object references:
- Format: $< ID_o, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, \ldots >$
- ID_o is a scene-stable object identifier; f_n is the time offset in seconds.
- x_{3d} and y_{3d} are LiDAR-derived global ground-plane coordinates in meters.
- Each tuple $(cam_{view}, x_{2d}, y_{2d})$ gives the 2D box center of the same object in a visible camera view.
- cam_{view} is one of north_image, south_image, east_image, or west_image.
- Multiple $(cam_{view}, x_{2d}, y_{2d})$ tuples may be included when the object is visible in multiple views.

LiDAR geometry:
- Coordinates are global, metric, and shared across frames.
- +X points south; +Y points east.
- Smaller X is farther north; larger X is farther south.
- Smaller Y is farther west; larger Y is farther east.

Figure 11: System prompt for Inter-3D VQA evaluation. For image-based VLMs, LiDAR-related instructions are removed to avoid unavailable-modality interference.

baselines, indicating that fine-grained dynamic attributes are still challenging. Since acceleration is sensitive to temporal estimation and can be noisy, it is difficult to infer reliably from limited LiDAR representations and short-term observations. These limitations suggest future improvements in tighter image–LiDAR alignment and stronger temporal motion modeling.

C.8 System Prompt

We design a dedicated system prompt for Inter-3D VQA to ensure consistent multimodal reasoning during evaluation. As shown in Fig. 11, the prompt specifies the available inputs, including four directional camera views and LiDAR-derived object and scene evidence. It further defines evidence usage rules: LiDAR evidence is prioritized for 3D geometry, global position, spatial layout, occlusion, and distance reasoning, while images are primarily used for visible appearance, signal states, and cross-view visual context. Note that the LiDAR evidence described in the prompt is used for grounding guidance. In Inter-Geo, the actual LiDAR information is provided as external feature tensors, aligned by DQNet, and injected into the decoder through LDA.

For image-based VLMs, we remove all LiDAR-related instructions from the prompt, including LiDAR evidence descriptions, coordinate-system definitions, and point-cloud-based geometric rules. This prevents unavailable modality information from interfering with image-only models, while keeping the image inputs, object references, and general reasoning constraints consistent across models.

To support grounded reasoning, the prompt also includes a unified object reference format containing scene-stable object IDs, temporal offsets, global coordinates, and multi-view 2D projections. In addition, it specifies the coordinate system and geometric rules for distance, lane, crosswalk, zone, and relative-direction reasoning when such information is available. We apply modality-consistent prompts across open-source, closed-source, and Inter-Geo models to ensure fair and controlled evaluation.

D Dataset Examples

D.1 Question Templates

In this section, we provide representative examples of question templates for each task category. Inter-3D VQA contains two complementary QA formats:

free-form QA and grounded QA. Figs. 12–15 illustrate representative free-form templates, while the corresponding grounded templates are shown in Figs. 16–19. Both formats cover the four task categories: Basic Perception, Spatial Reasoning, Temporal Reasoning, and Scene Understanding.

Within each category, templates are further organized into multiple subcategories, each targeting specific reasoning capabilities, such as object identification, geometric measurement, motion analysis, and interaction understanding. This design enables systematic coverage of diverse reasoning patterns in complex intersection scenarios.

During dataset construction, these placeholders are automatically instantiated using structured scene metadata extracted from synchronized multi-view images, LiDAR point clouds, object annotations, and HD maps. Specifically, `{ordinal}` represents object order in the same lane (e.g., first, second), `{object_type}` denotes object category (e.g., car, truck), `{lane_type}` indicates lane type (e.g., left-turn, right-turn), `{direction}` specifies approach direction (e.g., north, east), and `{spatial_relation}` encodes relative spatial relations (e.g., in front of, on the right, nearest). This process enables scalable and controllable QA generation while preserving semantic consistency across different reasoning tasks. By explicitly incorporating infrastructure topology and spatial relationships into template construction, Inter-3D VQA supports fine-grained multimodal reasoning beyond conventional image-centric VQA benchmarks.

D.2 Qualitative Evaluations

We visualize representative examples from the test set in Figs. 20 and 21. Each example includes the global point cloud, multi-view traffic images, the question, ground-truth answer, and predictions from Inter-Geo and several representative baseline models.

Figure 20 illustrates spatial and temporal reasoning cases in a dense daytime traffic scenario. In the spatial reasoning example, the target golf cart and car appear in different camera views, requiring cross-view association and distance estimation in the global 3D point cloud. Inter-Geo predicts a distance of 58.6 m, close to the ground truth of 56.9 m, while image-based baselines show larger errors. This demonstrates the benefit of explicit LiDAR geometry for distance estimation. The temporal reasoning example requires predicting the two-second future trajectory of a moving van. Inter-Geo better preserves the trajectory direction and displacement trend with only minor deviation from the ground truth, whereas other models show larger longitudinal or lateral drift. This suggests that aligned LiDAR representations provide effective geometric cues for trajectory-level spatial reasoning.

Figure 21 illustrates basic perception and scene understanding in a fast-moving nighttime scenario. In the basic perception case, nighttime illumination, motion blur, and occlusion make object counting challenging. Inter-Geo correctly identifies one truck in the intersection center, while other models overestimate the count due to misidentification of surrounding vehicles. In the scene understanding case, Inter-Geo detects the speeding-risk event and provides the relevant object category and speed estimate, whereas the baseline models fail to recognize the potential hazard. These examples show that global point clouds improve localization, motion estimation, and safety-critical reasoning.

Object Identify

Q: What is the type of the {ordinal} object in the {lane_type} lane on the {direction} approach?

Q: What is located {spatial_relation} the {ordinal} {object_type} in the {lane_type} lane on the {direction} approach?

Q: Does the {direction} approach contain any vulnerable road user?

Q: How many {object_type} are currently on the {direction} approach?

Appearance

Q: Which size best matches the {ordinal} {object_type} in the {lane_type} lane on the {direction} approach?

Q: How visible is the {ordinal} {object_type} in the {lane_type} lane on the {direction} approach?

Environment

Q: Please describe the current environmental conditions in the scene.

Q: Please describe the signal state for {lane_type} traffic on the {direction} approach.

Figure 12: Representative free-form question templates for Basic Perception.

Geometric Localization

Q: How far is the {ordinal} {object_type} in the {lane_type} lane on the {direction} approach from its relevant stop line?

Q: How far is the {spatial_relation} pedestrian on the {direction} crosswalk from the crossing exit area?

Q: What is the distance between the {ordinal_1} {object_type_1} in the {lane_type_1} lane on the {direction_1} approach and the {ordinal_2} {object_type_2} in the {lane_type_2} lane on the {direction_2} approach?

Q: Please identify the vehicle nearest to the {ordinal} {object_type} in the {lane_type} lane on the {direction} approach.

Topological Relation

Q: Which lane does the {ordinal} {object_type} in the {lane_type} lane on the {direction} approach currently occupy?

Q: Which area of the crosswalk is the {spatial_relation} pedestrian on the {direction} approach currently in?

Q: How many queued vehicles are currently in the {direction} {lane_type} lane?

Q: How many vehicles are currently within 5 m behind the {direction} stop line?

Q: Which lane currently has the longest queue right now?

Q: Is any vehicle blocking the crosswalk?

Figure 13: Representative free-form question templates for Spatial Reasoning.

Motion State

Q: Please describe the current motion state of the {ordinal} {object_type} in the {lane_type} lane on the {direction} approach.

Q: What maneuver is the {ordinal} {object_type} in the {lane_type} lane on the {direction} approach most likely executing?

Trajectory Prediction

Q: Please predict the short-term future trajectory of the {ordinal} {object_type} in the {lane_type} lane on the {direction} approach in 2 seconds.

Q: Which region is the {ordinal} {object_type} in the {lane_type} lane on the {direction} approach most likely to enter within the next 3 seconds?

Queue Dynamics

Q: Are straight-moving the {ordinal_1} {object_type_1} and the {ordinal_2} {object_type_2} in the {lane_type} lane on the {direction} approach maintaining a safe following gap?

Q: Which lane is most likely to form a long queue soon?

Multi-agent Conflict

Q: Is there a potential conflict between the {ordinal_1} {object_type_1} in the {lane_type_1} lane on the {direction_1} approach and the {ordinal_2} {object_type_2} in the {lane_type_2} lane on the {direction_2} approach right now?

Q: Which participant is most likely to conflict with the {ordinal} {object_type} in the {lane_type} lane on the {direction} approach?

Q: Please identify the key participant associated with the major potential safety hazards.

Q: What is the dominant conflict interaction pattern in the current scene?

Figure 14: Representative free-form question templates for Temporal Reasoning.

Intersection Status

Q: What is the overall traffic condition of the intersection?

Q: Please describe the motion status of traffic participants on the {direction} approach of the intersection.

Q: Provide a brief summary of the current intersection scene.

Q: Which approach to the intersection has the heaviest traffic?

Abnormal Event

Q: Is there a risk of speeding at the intersection?

Q: What is the most notable current abnormal event in the scene?

Traffic Suggestion

Q: How should traffic proceed through the intersection right now?

Q: What precaution should traffic take on the {direction} approach right now?

Q: How should traffic behave in the {lane_type} lane on the {direction} approach right now?

Q: What is the safest action for the {ordinal} {object_type} in the {lane_type} lane on the {direction} approach?

Figure 15: Representative free-form question templates for Scene Understanding.

Object Identify

Q: What is the type of $< o, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$?

Q: What is located {spatial_relation} of $< o, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$?

Q: Does the {direction} approach contain any vulnerable road user?

Q: How many {object_type} are currently on the {direction} approach?

Appearance

Q: Which size best matches $< o, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$?

Q: How visible is $< o, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$?

Environment

Q: Please describe the current environmental conditions in the scene.

Q: Please describe the signal state for {lane_type} traffic on the {direction} approach.

Figure 16: Representative grounded question templates for Basic Perception.

Geometric Localization

Q: How far is $< o, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$ from its relevant stop line?

Q: How far is $< o, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$ from the crossing exit area?

Q: What is the distance between $< o_1, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$ and $< o_2, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$?

Q: Please identify the vehicle nearest to $< o, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$.

Topological Relation

Q: Which lane does $< o, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$ currently occupy?

Q: Which area of the crosswalk is $< o, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$ currently in?

Q: How many queued vehicles are currently in the {direction} {lane_type} lane?

Q: How many vehicles are currently within 5 m behind the {direction} stop line?

Q: Which lane currently has the longest queue right now?

Q: Is any vehicle blocking the crosswalk?

Figure 17: Representative grounded question templates for Spatial Reasoning.

Motion State

Q: Please describe the current motion state of $< o, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$.

Q: What maneuver is $< o, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$ most likely executing?

Trajectory Prediction

Q: Please predict the short-term future trajectory of $< o, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$ in 2 seconds .

Q: Which region is $< o, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$ most likely to enter within the next 3 seconds?

Queue Dynamics

Q: Are straight-moving $< o_1, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$ and $< o_2, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$ maintaining a safe following gap?

Q: Which lane is most likely to form a long queue soon?

Multi-agent Conflict

Q: Is there a potential conflict between $< o_1, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$ and $< o_2, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$ right now?

Q: Which participant is most likely to conflict with $< o, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$?

Q: Please identify the key participant associated with the major potential safety hazards.

Q: What is the dominant conflict interaction pattern in the current scene?

Figure 18: Representative grounded question templates for Temporal Reasoning.

Intersection Status

Q: What is the overall traffic condition of the intersection?

Q: Please describe the motion status of traffic participants on the {direction} approach of the intersection.

Q: Provide a brief summary of the current intersection scene.

Q: Which approach to the intersection has the heaviest traffic?

Abnormal Event

Q: Is there a risk of speeding at the intersection?

Q: What is the most notable current abnormal event in the scene?

Traffic Suggestion

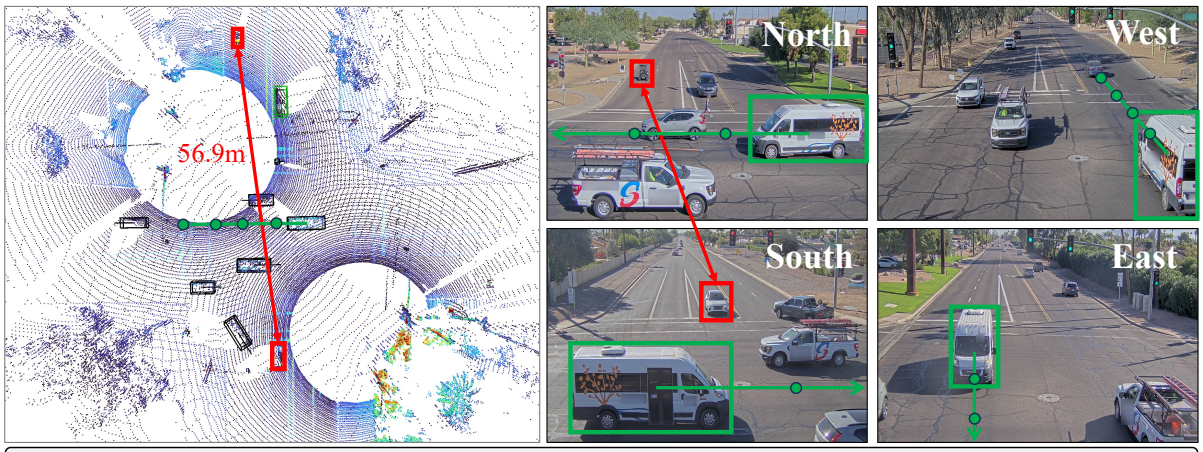
Q: How should traffic proceed through the intersection right now?

Q: What precaution should traffic take on the {direction} approach right now?

Q: How should traffic behave in the {lane_type} lane on the {direction} approach right now?

Q: What is the safest action for $< o, f_n, x_{3d}, y_{3d}, cam_{view}, x_{2d}, y_{2d}, ... >$?

Figure 19: Representative grounded question templates for Scene Understanding.



Task: Spatial Reasoning **QA Type:** Free-form

Question: What is the distance between the first golf cart in the right-turn lane on the north approach and the first car in the left-turn lane on the south approach?

Ground Truth: The golf cart is 56.9m from the car.

Inter-Geo: The golf cart is 58.6m from the car. ✓

Qwen3-VL-8B: The golf cart is 60.1m from the car. ✗

Senna-VLM: The golf cart is 53.0m from the car. ✗

TraffiX-Qwen: The golf cart is 43.7m from the car. ✗

Task: Temporal Reasoning **QA Type:** Grounded

Question: Please predict the short-term future trajectory of <o3, 2, 14.1, 17.7, north, 1536, 596, south, 606, 800, west, 1856, 570, east, 540, 596> in 2 seconds.

Ground Truth: Future trajectory: (0.0, -5.3), (-0.1, -10.6), (0.2, -16.4), (0.2, -22.5)

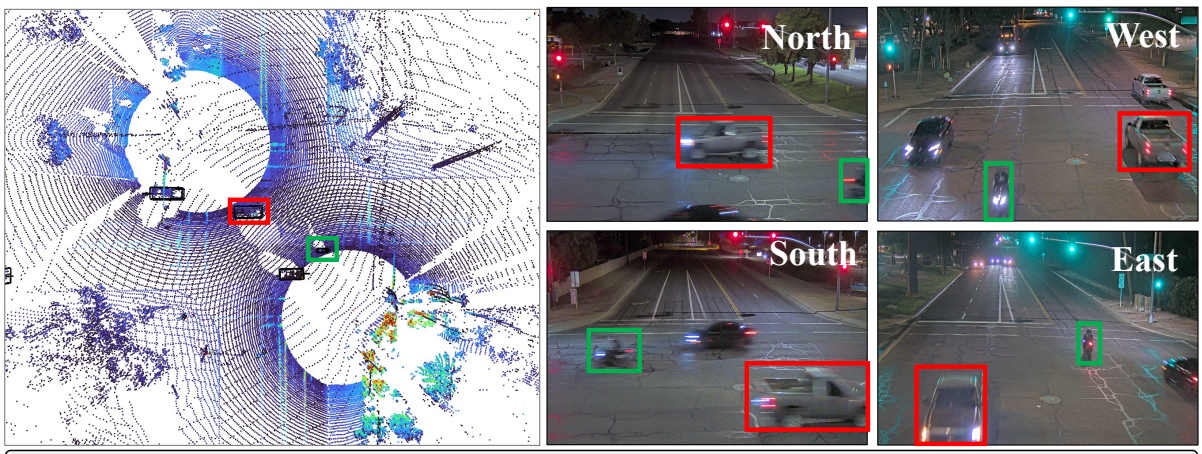
Inter-Geo: Future trajectory: (0.2, -7.3), (-0.1, -12.9), (0.6, -18.6), (0.6, -25.0) ✓

Qwen3-VL-8B: Future trajectory: (0.2, -14.7), (-0.6, -20.1), (0.4, -26.2), (0.9, -32.7) ✗

Senna-VLM: Future trajectory: (0.4, -14.5), (0.5, -21.0), (0.9, -27.5), (0.8, -35.3) ✗

TraffiX-Qwen: Future trajectory: (0.3, -15.9), (0.5, -22.1), (0.6, -28.7), (1.5, -37.0) ✗

Figure 20: Daytime examples of Spatial Reasoning and Temporal Reasoning.



Task: Basic Perception **QA Type:** Free-form

Question: How many trucks are currently in the center of the intersection?

Ground Truth: The center currently has 1 trucks.

Inter-Geo: The center currently has 1 trucks. ✓

Qwen3-VL-8B: The center currently has 2 trucks. ✗

Senna-VLM: The center currently has 3 trucks. ✗

TraffiX-Qwen: The center currently has 2 trucks. ✗

Task: Scene Understanding **QA Type:** Grounded

Question: Is there a risk of speeding at the intersection?

Ground Truth: Yes. A motorcycle in the center of the intersection is moving at about 21.4 m/s.

Inter-Geo: Yes. A motorcycle in the center of the intersection is moving at about 22.8 m/s. ✓

Qwen3-VL-8B: No. ✗

Senna-VLM: No. ✗

TraffiX-Qwen: No. ✗

Figure 21: Nighttime examples of Basic Perception and Scene Understanding.