BRIDGING PERCEPTION, BEHAVIOR, AND INTERACTION REASONING IN ROADSIDE TRAFFIC SCENES

Anonymous authors
Paper under double-blind review

ABSTRACT

Urban traffic scenes involve dynamic multi-agent behaviors and interactions, yet existing roadside benchmarks emphasize agent-level perception or holistic event reasoning, leaving behavior evolution and multi-agent interactions insufficiently modeled and evaluated. We introduce TRACE-Bench, a benchmark for roadside **T**raffic **R**easoning through **A**gent-**C**entric **E**volution. TRACE-Bench adopts a hierarchical annotation framework linking scene context, agent perception, behavior evolution, directed interactions, and causal reasoning. Continuous agent motion is decomposed into temporally grounded behavior units, while interactions are represented as directed influence–response relations. Built from real-world roadside RGB-Event recordings, TRACE-Bench comprises 201 finely annotated video sequences and 22,807 evidence-grounded QA pairs. We further propose an Event-Guided Sparse Multimodal framework that progressively reduces redundancy and fuses complementary RGB-Event information for efficient fine-grained traffic reasoning. Extensive experiments reveal limitations of existing MLLMs in behavior understanding and interaction reasoning, while demonstrating that our framework improves traffic perception and understanding with reduced inference latency.

1 INTRODUCTION

Urban traffic environments are highly dynamic and interactive, with heterogeneous road users, including vehicles and pedestrians, continuously influencing one another through shared road space and motion intentions Caesar et al. (2020); Geiger et al. (2013). Reliable understanding of such environments is essential for safety-critical applications, including intelligent transportation systems and autonomous driving. With their broad sensing coverage, roadside perception systems play a crucial role in traffic monitoring and management, increasingly requiring fine-grained spatiotemporal perception and reasoning to support safety analysis and decision-making Zhou et al. (2025a).

Recent multimodal large language models (MLLMs) Liu et al. (2023); Chen et al. (2024); Touvron et al. (2023) have demonstrated strong visual-textual modeling capabilities, enabling high-level reasoning and motivating their application to roadside traffic scenes Akter et al. (2025). Beyond fundamental perception, recent studies have introduced video question answering, scene description, and causal analysis, extending traffic understanding from agent-level perception to event-level semantic reasoning Guan et al. (2026); Vishal et al. (2026). However, existing datasets concentrate supervision primarily at the two ends of the reasoning chain: they characterize traffic agents in terms of category and location, while directly providing holistic event descriptions, causal explanations, or potential consequences. Agent behavior and multi-agent interaction, which connect these two ends, remain insufficiently modeled in a systematic and explicit manner and are rarely evaluated as standalone capabilities Ramanishka et al. (2018). In fact, agent behavior characterizes what a traffic participant does over time Tokic et al. (2026), while interaction relationships further describe who influences whom and how the affected agent responds Kung et al. (2024). Together, they form a critical intermediate layer that bridges low-level agent perception and high-level causal reasoning. Without such a structured reasoning chain, models may struggle to understand how complex traffic events emerge and evolve, while erroneous final predictions are also difficult to trace back to failures in agent perception, behavior recognition, interaction understanding, or causal inference.

To bridge this gap, we introduce TRACE-Bench, a benchmark for roadside **T**raffic **R**easoning through **A**gent-**C**entric **E**volution. To explicitly model agent behavior and interaction processes, it

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

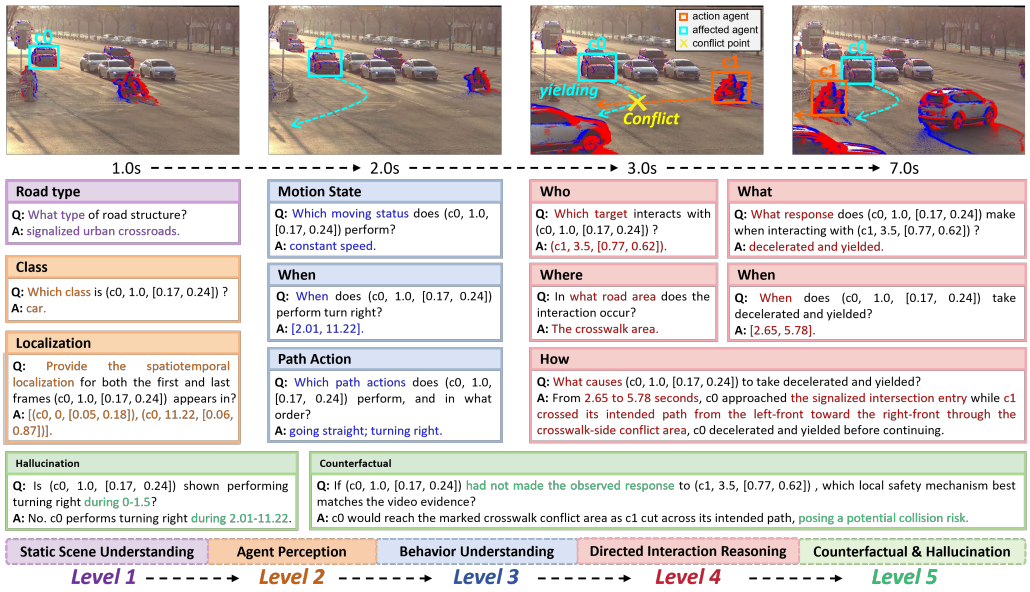


Figure 1: **An illustrative example of TRACE-Bench.** The benchmark covers five levels: **static scene understanding**, **agent perception**, **behavior understanding**, **directed interaction reasoning**, and **counterfactual & hallucination evaluation**. Agents are referenced by $(c, t, [x, y])$, where c is the persistent agent ID, t is the timestamp, and $[x, y]$ is the normalized bounding-box center, enabling precise spatiotemporal grounding in QA.

decomposes continuous agent motion into temporally grounded behavior units and further annotates directed influence–response relations between interacting agents. Built from real-world roadside RGB and event-camera recordings, TRACE-Bench covers diverse urban traffic scenarios and defines five task categories: static scene understanding, agent perception, behavior understanding, directed interaction reasoning, and counterfactual and hallucination evaluation, as shown in Fig. 1. The benchmark contains 201 finely annotated video sequences and 22,807 evidence-grounded QA pairs. Roadside traffic videos exhibit substantial spatiotemporal redundancy due to static backgrounds and relatively stable traffic states. To address this issue, we propose an Event-Guided Sparse Multimodal Framework that leverages event-derived dynamic cues to progressively select sparse, informative visual content and adaptively fuse complementary RGB-Event information for efficient fine-grained traffic reasoning. Our contribution can be summarized as follows:

(1) We introduce TRACE-Bench, a new roadside traffic scene understanding benchmark that provides a comprehensive evaluation pathway from perception to behavior understanding, interaction understanding, and causal reasoning.

(2) We propose an Event-Guided Sparse Multimodal framework that exploits dynamic cues captured by event cameras to guide the construction of low-redundancy and critical visual representations for efficient fine-grained traffic reasoning.

(3) We benchmark representative MLLMs to identify limitations in fine-grained traffic understanding and demonstrate the effectiveness of our framework in improving both reasoning performance and inference efficiency.

2 Related Work

2.1 Roadside Vision-Language Benchmarks

Infrastructure cameras provide continuous and global observations of intersections and road segments, while early roadside benchmarks primarily focused on object- and event-level perception. For example, CADP Shah et al. (2018) collected CCTV traffic videos to support object detection and accident prediction, representing the conventional paradigm of traffic event perception. With the

Table 1: Comparison of representative traffic understanding benchmarks relevant to roadside scenarios in terms of data characteristics and evaluation capabilities.

Dataset	Year	Viewpoint	Sensors	Spatiotemporal Grounding	Behavior Chain	Directed Interaction	Counter-factual	Halluci-nation	QA Scale
SUTD-TrafficQA (Xu et al., 2021)	2021	Mixed	RGB	✗	✗	✗	✓	✗	62.5K
TUMTraffic-VideoQA (Zhou et al., 2025b)	2025	Roadside	RGB	✓	✗	✗	✗	✗	85K
RoadSocial (Parikh et al., 2025)	2025	Mixed	RGB	✓	✗	✗	✓	✓	260K
MITS (Zhao et al., 2025)	2025	Roadside	RGB	✓	✗	✗	✗	✗	5M
RoadSafe365 (Liu et al., 2026)	2026	Mixed	RGB	✗	✗	✗	✗	✗	–
UDVideoQA (Vishal et al., 2026)	2026	Roadside	RGB	✓	✗	✗	✓	✗	28.8K
TRACE-Bench	2026	Roadside	**RGB + Event**	✓	✓	✓	✓	✓	22.8K

emergence of vision-language modeling, traffic benchmarks gradually shifted toward semantic understanding and reasoning. SUTD-TrafficQA Xu et al. (2021) introduced video question answering for event understanding, while WTS Kong et al. (2024) provided fine-grained, phase-level descriptions of pedestrian and vehicle states and contexts. However, these benchmarks primarily evaluate holistic event reasoning or dense description, without systematically disentangling agent-level perception from higher-level semantic understanding. TUMTraffic-VideoQA Zhou et al. (2025b) subsequently emphasized object-centric spatiotemporal understanding, but lacks high-level reasoning tasks. More recent benchmarks begin to jointly assess perception and reasoning: RoadSocial Parikh et al. (2025) incorporates temporal grounding and causal reasoning across diverse road videos, while UDVideoQA Vishal et al. (2026) targets multi-object spatiotemporal reasoning in complex urban scenes. However, as shown in Tab. 1, existing benchmarks still lack explicit evaluation of agent behavior and multi-agent interaction understanding, making it difficult to verify whether models truly capture how traffic events emerge and evolve. This limitation hinders both comprehensive understanding of complex traffic situations and fine-grained diagnosis of reasoning failures.

2.2 MLLMs for Traffic Scene Understanding

Recent MLLMs have extended traffic scene understanding beyond object recognition to visual reasoning. DriveLM Sima et al. (2024) employs graph-based QA to connect perception, prediction, and planning in driving scenarios. For fine-grained traffic video understanding, TrafficInternVL Phimsiri et al. (2025) combines structured fine-tuning with keyframe-based inputs, while TraffiX-Qwen Zhou et al. (2025b) investigates visual token sampling for roadside spatiotemporal QA. Beyond RGB-based models, EventDrive Lu et al. (2026) explores complementary RGB-Event sensing for driving tasks. Despite these advances, fine-grained traffic understanding requires modeling evolving behaviors and interactions from temporally extended observations, where substantial spatiotemporal redundancy poses challenges to efficient visual processing.

3 TRACE-Bench Dataset

TRACE-Bench is designed as a structured dataset for fine-grained roadside traffic understanding. It establishes a structured pathway from low-level perception to high-level reasoning while enabling independent evaluation of intermediate capabilities. As shown in Fig. 2, our dataset construction pipeline consists of four main stages: data acquisition and semantic clip selection, static scene and agent annotation, behavior and interaction annotation, and evidence-grounded QA construction.

3.1 Data Collection and Preprocessing

We collect RGB and event streams using a spatiotemporally synchronized RGB-Event camera deployed at fixed roadside locations. RGB camera provides rich appearance and scene-context cues, while the event camera captures stable and robust motion information by leveraging its high temporal resolution and high dynamic range Gallego et al. (2020); Yang et al. (2025). This complementary sensor configuration enables continuous and reliable observation of traffic agents and their interactions in highly dynamic traffic scenes with rapid motion and frequent illumination changes.

The dataset covers signalized and unsignalized intersections, merging and diverging areas, arterial road segments, and elevated road segments under diverse traffic densities and illumination conditions, as shown in Fig. 3(a). Continuous recordings are segmented according to the semantic

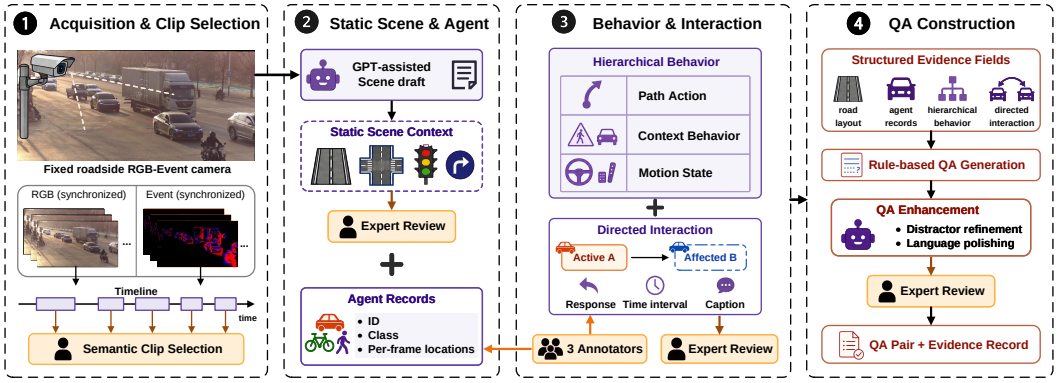


Figure 2: **Overview of the annotation and QA construction pipeline**, comprising four stages: semantic clip selection from synchronized RGB-Event data, static scene and agent annotation, hierarchical behavior and directed interaction annotation, and QA generation with quality control.

completeness of agent behaviors and interactions. Each clip lasts 10-35 seconds, retaining sufficient temporal context to observe behavior evolution and inter-agent interactions.

3.2 Agent-Centric Hierarchical Annotation

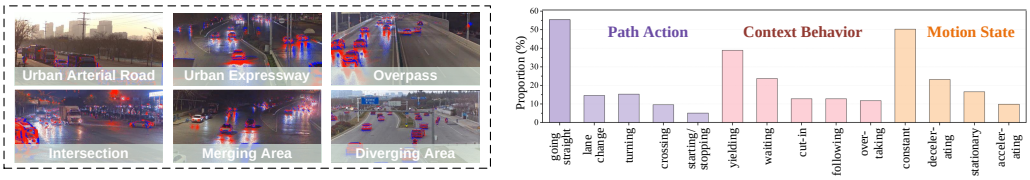
A central goal of TRACE-Bench is to explicitly model the intermediate processes connecting agent perception and event-level reasoning. To this end, we organize the annotations into progressively structured levels, from static scene context and persistent agent information to temporally grounded behaviors and directed interactions. We assigned three annotators with traffic-domain training and one traffic expert to review the annotation results.

Static scene and agent annotation. Each sequence is associated with a human-reviewed description of road layout, lane organization, traffic signs, and other static elements. Each visible agent is assigned a unique identity, category (car, e-bike, bicycle, motorcycle, tricycle, truck, bus, or pedestrian), and spatiotemporal locations. Persistent identities associate an agent with its complete visible history and provide the reference anchors for behavior and interaction questions.

Hierarchical behavior annotation. Traffic behavior is inherently compositional, and similar motions may correspond to different semantics depending on an agent's action within the surrounding traffic context. Inspired by the classic hierarchical theory of traffic behavior proposed by Michon (1985), we decompose continuous and complex agent motion along three complementary dimensions: path action, context behavior and motion state. This hierarchical decomposition provides a structured representation of traffic behavior across different semantic levels, with each behavior instantiated as a temporally grounded unit defined by an explicit semantic label and temporal interval.

Specifically, ***Path action*** describes goal-directed movements performed during navigation, including going straight, turning, lane changing, crossing, stopping, and starting. ***Context behavior*** captures responses conditioned on traffic context, including yielding, following, waiting, and overtaking, and is paired with a cause that specifies the relevant road or traffic condition. ***Motion state*** captures basic motion states, including constant, accelerating, decelerating, and stationary. Each behavior unit is temporally grounded with explicit start and end times, enabling TRACE-Bench to evaluate both agent behaviors and their temporal evolution. The behavior distribution is shown in Fig. 3(b), revealing the natural long-tailed distribution of behaviors in real-world roadside traffic.

Directed interaction annotation. An interaction is annotated only when one agent's behavior induces an observable response from another agent. Each record specifies the initiating agent, affected agent, response behavior, temporal interval, and a natural-language explanation of the influence–response process. Descriptions are drafted with GPT-5.5 Singh et al. (2025) assistance and then reviewed and refined by human experts. This formulation distinguishes directed behavioral influence from mere spatial proximity or co-occurrence. It identifies who affects whom, how the affected agent responds, and when the relation unfolds. To support fine-grained and temporally consistent annotation, we developed a dedicated tool for traffic behavior and interaction annotation (Fig. 10).



(a) Scene distribution **(b) Behavior statistics in different behavior layers**

Figure 3: Statistical distributions of the dataset.

3.3 QA Construction

Based on the hierarchical annotations, we construct an evidence-grounded QA benchmark in which each question is linked to explicit structured evidence, improving the traceability and verifiability of QA samples. In dense roadside traffic, agents with similar appearances and close proximity are difficult to distinguish through natural language alone. To address this ambiguity, we introduce an explicit spatiotemporal reference $(c, t, [x, y])$, where c denotes the unique agent ID, t the timestamp, and $[x, y]$ the normalized bounding-box center. This representation enables precise agent identification and provides a unified visual grounding interface for behavior, interaction, and spatiotemporal reasoning tasks. For QA generation, we manually design question templates (Fig. 11) and directly populate them with corresponding fields extracted from the structured annotations, ensuring strict consistency between the QA supervision and the underlying evidence.

TRACE-Bench contains five task categories. ***Static Scene Understanding*** evaluates road-type recognition, lane-number recognition, and topology description. ***Agent Perception*** covers category recognition and spatiotemporal grounding. ***Behavior Understanding*** evaluates behavior recognition, temporal localization, and behavior-chain understanding. ***Directed Interaction Reasoning*** evaluates interacting-agent identification, response-behavior understanding, temporal grounding and interaction-process description. Finally, ***Counterfactual and Hallucination Evaluation*** tests whether a model can reason about local changes to an observed interaction and reject premises that contradict the visual evidence. The detailed task hierarchy and descriptions are presented in Tab. 12.

In total, TRACE-Bench contains 201 sequences, approximately 1.4K finely annotated agents, and 22,807 evidence-grounded QA pairs. We split the data into training and validation subsets at a 7:3 ratio by collection scene, ensuring that the same physical location does not appear in both subsets. The split also preserves comparable distributions of task categories and question formats.

4 Method

4.1 Overview

Roadside videos exhibit substantial spatiotemporal redundancy. Temporally, adjacent frames share similar scene structures and traffic states. Spatially, fixed viewpoints repeatedly capture static backgrounds, and behavior-relevant information is concentrated in limited agents and local regions. By asynchronously recording pixel-level intensity changes, event cameras provide sparse dynamic cues that largely reflect agent motion and scene dynamics, offering a low-cost prior for spatiotemporal redundancy reduction. Based on this, we propose the Event-Guided Sparse Multimodal framework that progressively reduces redundancy while preserving RGB-Event information relevant to traffic reasoning, as shown in Fig. 4. First, **Event-Guided Frame Selection (EGFS)** performs temporal filtering using low-cost event statistics, while preserving low-motion context through uniform sampling. Next, **Adaptive Event Residual Fusion (AERF)** adaptively adjusts the contribution of event information to the fused features. Finally, **Dynamic-Relevance Token Selection (DRTS)** further removes spatial redundancy by jointly leveraging event activity and RGB semantic relevance.

4.1.1 Event-Guided Frame Selection (EGFS)

Fixed roadside videos exhibit substantial temporal redundancy, with adjacent frames sharing similar scene structures and agent states. In contrast, fine-grained traffic reasoning relies on dynamic changes in agent behaviors and interactions. Event cameras capture such changes at high temporal

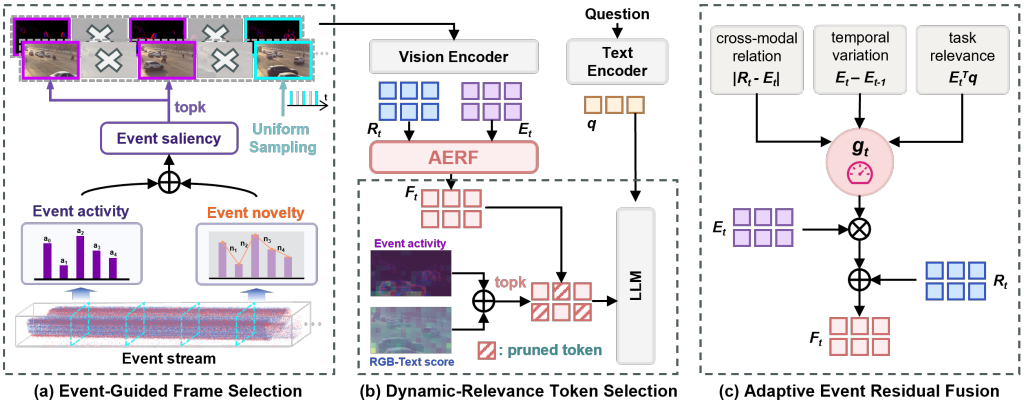


Figure 4: **Overview of the Event-Guided Sparse Multimodal Framework.** (a) EGFS reduces temporal redundancy while preserving context. (b) DRTS removes spatial redundancy based on event activity and RGB semantic relevance. (c) AERF dynamically integrates complementary event information into RGB features.

resolution, providing fine-grained motion cues. Therefore, we leverage event-based motion cues as a temporal routing signal to identify motion-salient frames based on **event activity** and **temporal novelty**, while incorporating uniform sampling to preserve temporal context.

Specifically, given an asynchronous event stream of candidate frame t, each event is represented as $e_k = (x_k, y_k, \tau_k, p_k)$, where (x_k, y_k), τ_k, and p_k denote the spatial location, timestamp, and polarity of the k-th event, respectively. For each candidate frame t, computing event histogram:

$$\mathbf{H}_t(u, v) = \sum_{e_k} \mathbb{I}\left[(x_k, y_k) = (u, v)\right]. \tag{1}$$

This representation compresses the asynchronous event stream into a spatial activity distribution while retaining the overall event response. Based on $\mathbf{H}_t$, we define the event activity score that jointly captures response magnitude and spatial coverage:

$$a_t = \log\left(1 + \|\mathbf{H}_t\|_1\right) \cdot \frac{1}{HW} \sum_{u=1}^{H} \sum_{v=1}^{W} \mathbb{I}\left[\mathbf{H}_t(u, v) > 0\right]. \tag{2}$$

The first term captures the overall event response, with logarithmic compression mitigating the dominance of extremely high event counts. The second measures the spatial coverage of event activity, favoring responses distributed across broader regions.

Event activity captures the magnitude and spatial extent of event responses but does not reflect temporal changes in motion patterns. We therefore introduce a temporal novelty score by comparing normalized event histograms of adjacent frames:

$$n_t = \left\| \frac{\mathbf{H}_t}{\|\mathbf{H}_t\|_1} - \frac{\mathbf{H}_{t-1}}{\|\mathbf{H}_{t-1}\|_1} \right\|_1. \tag{3}$$

This score quantifies changes in the spatial distribution of event responses, potentially reflecting motion-state transitions. Finally, the event saliency score is computed as:

$$s_t^{\text{event}} = \frac{1}{2} a_t + \frac{1}{2} n_t. \tag{4}$$

Given the temporal retention ratio k_1, we combine event saliency-based selection with uniform sampling. The former prioritizes dynamically informative observations, while the latter preserves temporal coverage and contextual information across the video.

4.1.2 Adaptive Event Residual Fusion (AERF)

RGB and Event provide complementary but asymmetric information for roadside traffic understanding. RGB offers rich semantic cues, while event data captures fine-grained dynamics for behavior

transitions and interactions but has limited semantic representation due to sparse appearance information and low signal-to-noise ratio. Therefore, we treat event features as complementary dynamic residuals and propose **Adaptive Event Residual Fusion (AERF)** to enhance RGB representations. AERF employs a gating mechanism conditioned on cross-modal relations, temporal dynamics, and task relevance to adaptively regulate event contributions across frames and spatial locations.

To estimate the contribution of event features, we construct the fusion gate using three complementary cues. **First**, cross-modal relations are characterized by the difference $|\mathbf{R}_t - \mathbf{E}_t|$, where $\mathbf{R}_t$ and $\mathbf{E}_t$ denote the local RGB and event representations at frame t, respectively. This difference exposes local representation discrepancies that provide complementary information for fusion. **Second**, temporal dynamics are captured by $\mathbf{E}_t - \mathbf{E}_{t-1}$, which characterize adjacent-frame variations that accompany motion-state transitions or evolving interactions. **Finally**, we introduce event-task relevance $\mathbf{E}_t^\top \mathbf{q}$ as a task-conditioned cue, where $\mathbf{q}$ denotes the question representation, to account for varying demands on event information across tasks. The fusion process is defined as

$$g_t = \sigma\left(f_g\left[|\mathbf{R}_t - \mathbf{E}_t|\,; \mathbf{E}_t^\top \mathbf{q}; \mathbf{E}_t - \mathbf{E}_{t-1}\right]\right), \tag{5}$$

$$\hat{\mathbf{R}}_t = \mathbf{R}_t + g_t \mathbf{E}_t, \tag{6}$$

where f_g denotes a two-layer MLP, $[;]$ is concatenation. By jointly modeling cross-modal relations, temporal variation and task relevance, the gating mechanism adaptively modulates the contribution of event information to the fused feature.

4.1.3 Dynamic-Relevance Token Selection (DRTS)

To further reduce spatial redundancy, we propose **Dynamic-Relevance Token Selection (DRTS)**, which jointly leverages event-based dynamic priors and RGB-Text semantic relevance for fine-grained token selection. Specifically, event activity highlights regions with strong motion responses, while RGB-Text relevance identifies regions semantically aligned with the current task. The spatial importance score is computed as:

$$s_t^{spatial} = \frac{1}{2}\mathbf{H}_t + \frac{1}{2} r_t, \tag{7}$$

where $\mathbf{H}_t$ is the event histogram map, r_t is computed based on the similarity between the question tokens and RGB tokens. Given the spatial retention ratio k_2, we retain the top-k_2 tokens.

5 Experiments

5.1 Experiment Details

Implementation Details. We build our framework on Qwen3-VL-4B Bai et al. (2025) to balance accuracy and efficiency, with RGB and event modalities sharing a visual encoder. Videos are sampled at 2 FPS, and event streams are accumulated over intervals between consecutive frames of the original RGB video and converted into event images. Both temporal and spatial retention ratios are set to 50%. During supervised fine-tuning, we freeze the visual encoder and language model, optimizing only the vision–language projector, fusion module, and attention-projection LoRA Hu et al. (2021) adapters ($r = 16$, $\alpha = 32$, dropout $= 0.05$). Training runs for one epoch using AdamW with a learning rate of $1 \times 10^{-}5$ and a cosine learning rate schedule.

Evaluation Metrics. We evaluate multiple-choice QA (MCQ) using accuracy, temporal grounding using MAE (T-MAE), and spatiotemporal localization using Spatial L1 (S-L1) and Spatiotemporal Error (ST Err.). Open-ended QA is evaluated using BERTScore-F1 Zhang et al. (2019), BLEU-4 Papineni et al. (2002), ROUGE-L Lin (2004), and CIDEr Vedantam et al. (2015). Metric definitions are detailed in Appendix A.1.1.

Evaluated Models. We evaluate representative open-source MLLMs in a zero-shot setting, including VideoLLaMA3-2B/7B Zhang et al. (2025), InternVL3.5-2B/4B/8B Wang et al. (2025), and Qwen3-VL-2B/4B/8B. We further fine-tune Qwen3-VL-4B with RGB-only and Event-only inputs and include two domain-specific models: TUMTraffic-Qwen-7B Zhou et al. (2025b) for RGB-based traffic understanding and EventGPT-7B Liu et al. (2025) for event-based multimodal understanding. We also compare RGB-Event fusion strategies by replacing our fusion module with element-wise addition (AD) and cross-attention (CA) Vaswani et al. (2017).

Table 2: Performance comparison of zero-shot models and fine-tuned baselines on TRACE-Bench. T-MAE, S-L1, and ST Err. measure temporal, spatial, and spatiotemporal grounding, respectively.

	Model	Acc. ↑	T-MAE ↓	S-L1 ↓	ST Err. ↓	BERT-F1 ↑	ROUGE-L ↑	BLEU-4 ↑	CIDEr ↑
Open-source Models									
RGB	VideoLLaMA3-2B	55.00	5.72	0.3939	0.5652	0.1699	0.1907	0.0473	0.0409
RGB	VideoLLaMA3-7B	62.16	5.10	0.3743	0.5436	0.1786	0.2342	0.0469	0.0726
RGB	InternVL3.5-2B	58.75	4.71	0.3460	0.5664	0.1419	0.2103	0.0353	**0.0925**
RGB	InternVL3.5-4B	**63.75**	4.43	0.4248	0.5622	**0.2134**	**0.2505**	0.0438	0.0555
RGB	InternVL3.5-8B	53.93	**4.40**	**0.2342**	0.5577	0.1716	0.1965	**0.0475**	0.0562
RGB	Qwen3-VL-2B	48.64	9.28	0.3604	0.5450	0.1638	0.2403	0.0264	0.0576
RGB	Qwen3-VL-4B	63.45	4.64	0.4137	**0.5391**	0.2066	0.2277	0.0430	0.0529
RGB	Qwen3-VL-8B	58.30	**4.40**	0.2657	0.5534	0.2096	0.2273	0.0337	0.0295
Fine-tuned Models									
RGB	Qwen3-VL-RGB-4B	78.13	4.46	0.1363	0.5132	0.2666	0.3170	0.0466	0.7663
RGB	TUMTraffic-Qwen-7B	65.13	4.61	0.1354	0.5484	0.2689	0.3210	0.0343	0.7796
Event	Qwen3-VL-Event-4B	76.24	4.36	0.1361	0.5117	0.2626	0.3237	0.0411	0.8112
Event	Event-GPT-7B	69.29	5.38	0.2756	0.5197	0.1792	0.2648	0.0255	0.5317
RGB-Event	RGB-Event Add	73.91	4.51	0.2015	0.5025	0.1775	0.2555	0.0300	0.0933
RGB-Event	RGB-Event CA	77.60	4.31	0.1373	0.5135	0.2708	0.3232	**0.0490**	0.7678
RGB-Event	**Ours**	**78.77**	**4.25**	**0.1305**	**0.5009**	**0.2752**	**0.3281**	0.0489	**0.8154**

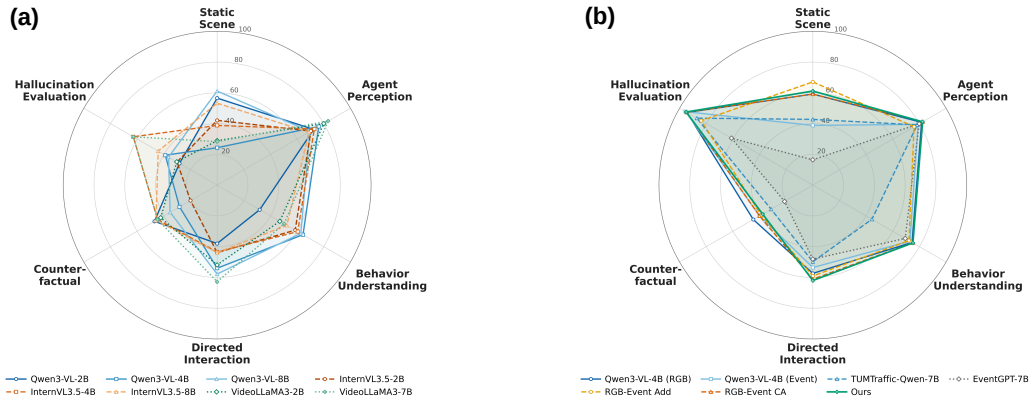


Figure 5: Radar charts of zero-shot and fine-tuned performance on the MCQ task, respectively.

5.2 Main Results

Tab. 2 presents the overall performance across different task types, while Fig. 5 visualizes MCQ accuracy across different levels of traffic understanding.

Zero-shot Performance. No open-source MLLM consistently excels across all capabilities. For instance, InternVL3.5-4B achieves the highest accuracy (63.75%), BERT-F1 (0.2134), and ROUGE-L (0.2505), but exhibits the largest spatial localization error (S-L1: 0.4248). This demonstrates the diagnostic value of TRACE-Bench in revealing uneven model capabilities across different dimensions of traffic understanding. As shown in Fig. 5(a), existing MLLMs exhibit substantial capability gaps across different levels of traffic understanding. Average accuracy declines from 74.24% in agent perception to 53.46% in behavior understanding and 49.47% in directed interaction reasoning, with further declines in counterfactual (38.53%) and hallucination (41.78%) evaluation. Moreover, model scaling yields inconsistent improvements across task levels, indicating that these capability gaps are not simply resolved by increasing model size. These findings indicate that strong agent perception does not necessarily translate into understanding behavior evolution and interaction relationships, demonstrating the diagnostic value of TRACE-Bench in identifying capability bottlenecks.

Fine-tuned Performance. Fine-tuning substantially improves performance. For instance, Qwen3-VL-4B-RGB improves accuracy from 63.45% to 78.13% while reducing S-L1 from 0.4137 to 0.1363. However, the gap between agent perception and interaction reasoning persists (Fig. 5(b)), indicating that higher-level reasoning remains challenging even with task-specific supervision. Modality comparisons reveal complementary strengths. RGB-only Qwen3-VL-4B substantially outper-

Table 3: Efficiency comparison.

Model	Latency(s)	TFLOPs	Memory(G)
Qwen3-VL-4B	2.34	195.69	10.44
RGB-Event Add	2.03	77.33	**9.21**
RGB-Event CA	1.91	77.32	9.88
Ours	**1.82**	**77.30**	**9.21**

Table 4: Performance of our components.

EGFS	AERF	DRTS	Acc. ↑	T-MAE ↓	ST Err. ↓	BERT-F1 ↑	Latency ↓
RGB-Event Add*			74.16	4.39	0.5183	0.1841	3.95
	✓		78.08	4.35	0.5124	0.2641	4.00
✓	✓		78.16	**4.22**	**0.4975**	0.2725	1.94
	✓	✓	78.55	4.32	0.5140	0.2623	3.10
✓	✓	✓	**78.77**	4.25	0.5009	**0.2752**	**1.82**

Table 5: Ablation of event saliency based sampling and uniform sampling in temporal selection.

event saliency	uniform sampling	Acc. ↑	T-MAE ↓	ST Err. ↓	BERT-F1 ↑
✓		78.44	4.28	**0.4992**	0.2736
	✓	77.31	4.43	0.5180	0.2720
0.5	0.5	78.61	4.39	0.5120	**0.2757**
0.7	0.3	78.65	4.37	0.5051	0.2678
0.85	**0.15**	**78.77**	**4.25**	0.5009	0.2752
0.9	0.1	78.58	4.29	0.5011	0.2741

Table 6: Ablation of event activity ($\mathbf{H}_t$) and RGB-Text relevance (r_t) in spatial token selection.

$\mathbf{H}_t$	r_t	Acc. ↑	T-MAE ↓	ST Err. ↓	BERT-F1 ↑
✓		78.59	**4.24**	0.5061	0.2729
	✓	78.23	4.32	0.5138	0.2704
✓	✓	**78.77**	4.25	**0.5009**	**0.2752**

forms Event-only in static-scene understanding (59.21% vs. 38.82%), whereas Event-only achieves lower T-MAE (4.36 vs. 4.46) and higher CIDEr (0.8112 vs. 0.7663), reflecting the complementary roles of RGB appearance semantics and event-based dynamic cues. However, naive RGB-Event addition reduces overall accuracy, indicating that effective fusion is essential to exploit cross-modal complementarity. Through adaptive residual fusion and spatiotemporal token selection, our method achieves the best overall performance, improving behavior understanding and directed interaction reasoning by 0.67 and 4.65 percentage points, respectively, over RGB-only Qwen3-VL-4B. Meanwhile, it reduces latency, TFLOPs, and memory usage by 22.2%, 60.5%, and 11.8%, respectively (Tab. 3), demonstrating improvements in both accuracy and efficiency. Furthermore, the evidence-linked QA design of TRACE-Bench supports cross-level error attribution analysis, revealing model-specific understanding bottlenecks (see Appendix A.2.1).

5.3 Ablation Study

Component Ablation. As shown in Tab. 4, we adopt RGB-Event Add*, which directly combines RGB and event features through addition without spatiotemporal sparsification, as the baseline. AERF improves accuracy and BERT-F1 by adaptively modulating event contributions. Incorporating EGFS and DRTS further improves performance while reducing inference latency. The full model achieves the best overall performance, while reducing latency from 3.95 to 1.82.

Temporal Sampling Strategy. As shown in Tab. 5, event-guided selection outperforms uniform sampling, demonstrating the benefit of prioritizing dynamically informative frames. Combining both strategies further improves performance, with the 0.85/0.15 ratio achieving the best overall performance. We therefore adopt this configuration in all experiments. These results suggest that uniform sampling can complement event-guided selection by improving temporal context coverage.

Spatial Selection Cues. As shown in Tab. 6, event activity alone outperforms RGB-Text relevance, while combining both cues further improves performance, demonstrating the complementarity of dynamic saliency and task relevance in guiding spatial token selection.

6 Conclusion

We introduce TRACE-Bench, a comprehensive traffic understanding benchmark spanning static-scene understanding, agent perception, behavior understanding, directed interaction reasoning, and causal reasoning. It provides a structured evaluation of models' ability to understand traffic behaviors and event evolution. We further propose an Event-Guided Sparse Multimodal Framework that selects and fuses informative visual cues for fine-grained traffic reasoning, balancing understanding accuracy and inference efficiency. Extensive experiments reveal substantial gaps between low-level perception and higher-level behavior and interaction understanding in existing MLLMs, while demonstrating that our framework improves traffic understanding and spatiotemporal grounding performance with reduced inference latency.

AI USE STATEMENT

We used generative AI tools to assist with manuscript polishing and dataset construction, including drafting scene and interaction annotations and refining question-answer pairs. All AI-assisted annotations and QA content were reviewed and corrected by human annotators against visual evidence and structured annotations.

ETHICS STATEMENT

This study aims to advance fine-grained roadside traffic scene understanding and multimodal model evaluation. TRACE-Bench is constructed from real-world roadside RGB-Event recordings, with appropriate privacy protection measures applied to potentially identifiable information about pedestrians and vehicles. For annotations and question-answer pairs assisted by generative AI, we conduct human review and verification against visual evidence to improve annotation reliability and prevent factual errors, bias, and inappropriate content.

REFERENCES

Sanjeda Akter, Ibne Farabi Shihab, and Anuj Sharma. Large language models for crash detection in video: A survey of methods, datasets, and challenges. *arXiv preprint arXiv:2507.02074*, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11618–11628. IEEE, 2020.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(1):154–180, 2020.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.

Runwei Guan, Rongsheng Hu, Shangshu Chen, Ningyuan Xiao, Xue Xia, Jiayang Liu, Beibei Chen, Ziren Tang, Ningwei Ouyang, Shaofeng Liang, et al. Roadscenevqa: Benchmarking visual question answering in roadside perception systems for intelligent transportation system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4366–4375, 2026.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Quan Kong, Yuki Kawana, Rajat Saini, Ashutosh Kumar, Jingjing Pan, Ta Gu, Yohei Ozao, Balazs Opra, Yoichi Sato, and Norimasa Kobori. Wts: A pedestrian-centric traffic video dataset for fine-grained spatial-temporal understanding. In *European Conference on Computer Vision*, pp. 1–18. Springer, 2024.

Chi-Hsi Kung, Chieh-Chi Yang, Pang-Yuan Pao, Shu-Wei Lu, Pin-Lun Chen, Hsin-Cheng Lu, and Yi-Ting Chen. Riskbench: A scenario-based benchmark for risk identification. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14800–14807. IEEE, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Shaoyu Liu, Jianing Li, Guanghui Zhao, Yunjian Zhang, Xin Meng, Fei Richard Yu, Xiangyang Ji, and Ming Li. Eventgpt: Event stream understanding with multimodal large language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29139–29149. IEEE, 2025.

Xinyu Liu, Darryl C Jacob, Yuxin Liu, Xinsong Du, Muchao Ye, Bolei Zhou, and Pan He. Understanding real-world traffic safety through roadsafe365 benchmark. *arXiv preprint arXiv:2602.07212*, 2026.

Dongyue Lu, Rong Li, Ao Liang, Lingdong Kong, Wei Yin, Lai Xing Ng, Benoit R Cottereau, Camille Simon Chane, and Wei Tsang Ooi. Eventdrive: Event cameras for vision-language driving intelligence. *arXiv preprint arXiv:2606.18242*, 2026.

John A Michon. A critical view of driver behavior models: what do we know, what should we do? In *Human behavior and traffic safety*, pp. 485–524. Springer, 1985.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Chirag Parikh, Deepti Rawat, Tathagata Ghosh, Ravi Kiran Sarvadevabhatla, et al. Roadsocial: A diverse videoqa dataset and benchmark for road event understanding from social video narratives. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19002–19011. IEEE, 2025.

Sasin Phimsiri, Sarut Sunpawatr, Riu Cherdchusakulchai, Pornprom Kiawjak, Teepakorn Tosawadi, Suchat Tungjitnob, Visarut Trairattanapa, Supawit Vatathanavaro, Wasu Kudisthalert, Chaitat Utintu, et al. Trafficinternvl: Spatially-guided fine-tuning with caption refinement for fine-grained traffic safety captioning and visual question answering. In *2025 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 5358–5365. IEEE, 2025.

Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 7699–7707. IEEE, 2018.

Ankit Parag Shah, Jean-Bapstite Lamare, Tuan Nguyen-Anh, and Alexander Hauptmann. Cadp: A novel dataset for cctv traffic camera based accident analysis. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–9. IEEE, 2018.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pp. 256–274. Springer, 2024.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Boris Tokic, Constantin Selzer, and Fabian B Flohr. Commandlm: Data driven behavior level descriptor for ego vehicles. In *2026 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1414–1420. IEEE, 2026.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Joseph Raj Vishal, Nagasiri Poluri, Katha Naik, Rutuja Patil, Kashyap Hegde Kota, Krishna Vinod, Prithvi Jai Ramesh, Mohammad Farhadi, Yezhou Yang, and Bharatesh Chakravarthi. Udvideoqa: A traffic video question answering dataset for multi-object spatio-temporal reasoning in urban dynamics. *arXiv preprint arXiv:2602.21137*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Li Xu, He Huang, and Jun Liu. Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9873–9883. IEEE, 2021.

Nan Yang, Yang Wang, Zhanwen Liu, Meng Li, Yisheng An, and Xiangmo Zhao. Smamba: Sparse mamba for event-based object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 9229–9237, 2025.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Kaikai Zhao, Zhaoxiang Liu, Peng Wang, Xin Wang, Zhicheng Ma, Yajun Xu, Wenjing Zhang, Yibing Nan, Kai Wang, and Shiguo Lian. Mits: A large-scale multimodal benchmark dataset for intelligent traffic surveillance. *Image and Vision Computing*, pp. 105736, 2025.

Wei Zhou, Li Yang, Lei Zhao, Runyu Zhang, Yifan Cui, Hongpu Huang, Kun Qie, and Chen Wang. Vision technologies with applications in traffic surveillance systems: A holistic survey. *ACM Computing Surveys*, 58(3):1–47, 2025a.

Xingcheng Zhou, Konstantinos Larintzakis, Hao Guo, Walter Zimmer, Mingyu Liu, Hu Cao, Jiajie Zhang, Venkatnarayanan Lakshminarasimhan, Leah Strand, and Alois C Knoll. Tumtraffic-videoqa: A benchmark for unified spatio-temporal video understanding in traffic scenes. *arXiv preprint arXiv:2502.02449*, 2025b.

System Prompt:
You are an AI assistant specializing in the analysis of roadside traffic scenes from surveillance videos. Each object's spatiotemporal reference is represented as [(c, t_start, [x_start, y_start]), (c, t_end, [x_end, y_end])] or (c, t, [x, y]), where c is the unique anonymized object identifier, t is the timestamp in seconds, and x and y are normalized center-point coordinates between 0 and 1. Answer each question based only on the visible video content. Follow the requested answer format exactly and avoid adding extra explanation unless the question asks for it.

Figure 6: The system prompt used in the experiments.

Table 7: Cross-level diagnostic analysis of fine-tuned models. P→B Error denotes the path-action recognition error conditioned on correct agent-category recognition. B→I Error denotes the categorical interaction-event recognition error conditioned on correct path-action recognition.

Model	P→B Error ↓	B→I Error ↓
Qwen3-VL-4B-RGB	24.10	33.33
Qwen3-VL-4B-Event	25.58	37.80
Ours	**23.06**	**29.35**

A Appendix

This appendix provides additional details on TRACE-Bench and our method. We first present the evaluation metrics and system prompt used in experiments (A.1). We then provide cross-level diagnostic analysis (A.2.1), comparisons under different illumination conditions (A.2.2), qualitative comparisons (A.2.3), additional ablation studies (A.2.4), and visualizations of spatiotemporal token selection (A.2.5). Finally, we provide further details on the dataset construction process (A.3).

A.1 More Experiment Details

A.1.1 Evaluation Metrics

TRACE-Bench contains multiple-choice, temporal grounding, spatiotemporal localization, and open-ended questions. We adopt evaluation metrics according to the required answer format and report the mean score over all samples of each question type.

Multiple-choice questions. Multiple-choice questions cover most recognition and reasoning tasks in TRACE-Bench. We extract the predicted option label and report Accuracy (%).

Temporal grounding questions. Temporal grounding requires predicting the start and end times of a behavior or interaction. Let $[t_s, t_e]$ and $[\hat{t}_s, \hat{t}_e]$ denote the ground-truth (GT) and predicted temporal intervals, respectively. We report the mean absolute boundary error:

$$\mathrm{MAE} = \frac{1}{2}\left(|t_s - \hat{t}_s| + |t_e - \hat{t}_e|\right). \tag{8}$$

Spatiotemporal localization questions. We evaluate spatial localization using the normalized center coordinates of the object. For a GT point (x, y) and prediction $(\hat{x}, \hat{y})$, we report:

$$\text{Spatial L1} = |x - \hat{x}| + |y - \hat{y}|. \tag{9}$$

For spatiotemporal localization, the model predicts the first and last observed states of an object, each represented by an agent ID, timestamp, and spatial location. We report:

$$\text{ST Error} = \frac{1}{2} \sum_{k \in \text{first,last}} \sqrt{\left(\frac{|t_k - \hat{t}_k|}{D}\right)^2 + (x_k - \hat{x}_k)^2 + (y_k - \hat{y}_k)^2}, \tag{10}$$

where D is the video duration.

Open-ended questions. For open-ended responses, we use BERTScore-F1 to measure semantic similarity, together with BLEU-4, ROUGE-L, and CIDEr as complementary text-generation metrics. Higher scores indicate better agreement with the reference answers.

Table 8: Performance of fine-tuned models on daytime sequences.

Model	Acc. ↑	T-MAE ↓	S-L1 ↓	ST Err. ↓	BERT-F1 ↑	ROUGE-L ↑	BLEU-4 ↑	CIDEr ↑
Qwen3-VL-4B-RGB	71.59	5.30	0.1177	0.4420	0.3016	0.3238	0.0525	0.7350
Qwen3-VL-4B-Event	68.93	5.15	0.1261	0.4351	0.2791	0.3323	0.0362	0.8019
Ours	**71.96**	**4.91**	**0.1133**	**0.4217**	**0.3047**	**0.3396**	**0.0581**	**0.8049**

Table 9: Performance of fine-tuned models on nighttime sequences.

Model	Acc. ↑	T-MAE ↓	S-L1 ↓	ST Err. ↓	BERT-F1 ↑	ROUGE-L ↑	BLEU-4 ↑	CIDEr ↑
Qwen3-VL-4B-RGB	82.25	3.87	0.1392	0.5485	0.2404	0.3107	0.0431	0.8874
Qwen3-VL-4B-Event	80.84	3.80	0.1410	0.5498	**0.2457**	0.3147	**0.0466**	0.8950
Ours	**83.05**	**3.79**	**0.1390**	**0.5402**	0.2449	**0.3163**	0.0436	**0.8959**

A.1.2 System Prompt

We design a dedicated system prompt for experiments on our dataset, as shown in Fig. 6. The same prompt is applied consistently to all experiments to ensure a fair and standardized comparison across different models.

A.2 More Experiment Results

A.2.1 Cross-Level Diagnostic Analysis

To determine whether model reasoning failures stem from insufficient low-level perception or high-level understanding capabilities, we conduct a cross-level diagnostic analysis using linked QA pairs associated with the same agents and interaction instances. Specifically, we measure the behavior recognition error conditioned on correct agent-category recognition (P→B Error) and the interaction understanding error conditioned on correct path-action recognition (B→I Error). Unlike conventional task-wise accuracy, these conditional error rates reveal whether higher-level understanding failures persist even when the associated lower-level recognition tasks are correctly completed.

As shown in Tab. 7, RGB-only Qwen3-VL-4B exhibits a P→B Error of 24.10%, indicating that nearly one-quarter of the linked behavior questions remain incorrectly answered despite correct agent-category recognition. More notably, its B→I Error reaches 33.33%, revealing that correct recognition of individual path actions does not necessarily translate into an accurate understanding of the associated interactions. Similar patterns are observed in the Event-only model, with conditional error rates of 25.58% and 37.80%, respectively. These findings highlight the limitations of

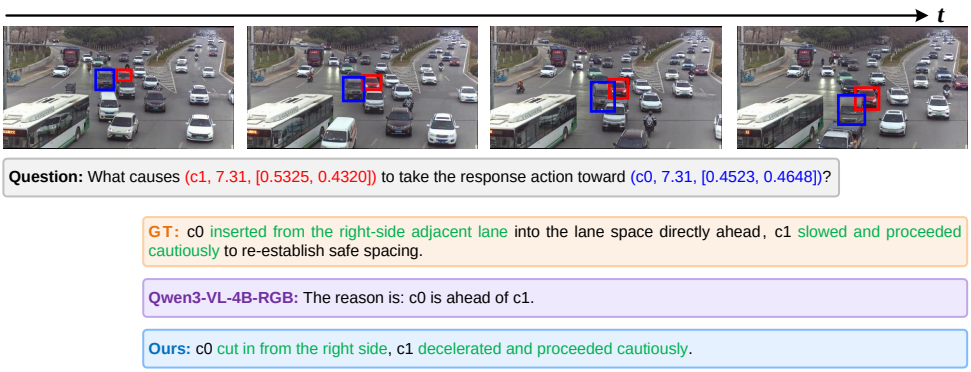


Figure 7: Qualitative comparison of interaction-process descriptions in a daytime roadside scene. The initiating and responding agents are highlighted in blue and red, respectively. Our method identifies the cut-in maneuver and the subsequent deceleration, whereas the RGB-only Qwen3-VL-4B mainly describes the relative positions of the two agents.

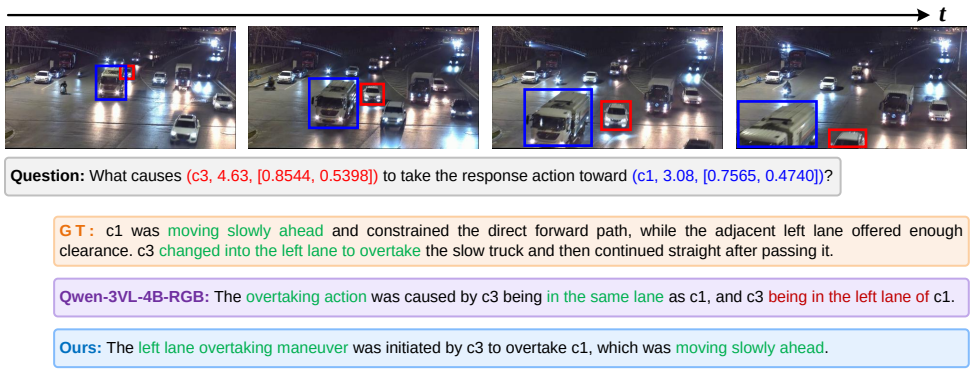


Figure 8: Qualitative comparison of interaction-process descriptions in a nighttime roadside scene. The initiating and responding agents are highlighted in blue and red, respectively. Compared with the RGB-only Qwen3-VL-4B, our method better identifies the slow leading vehicle as the reason for the subsequent left-lane overtaking maneuver.

Table 10: Ablation of input frame sampling rates (FPS).

FPS	Acc. ↑	T-MAE ↓	S-L1 ↓	ST Err. ↓	BERT-F1 ↑	ROUGE-L ↑	BLEU-4 ↑	CIDEr ↑	Latency ↓
without	33.17	-	-	-	-	-	-	-	-
1	78.41	4.30	0.1332	0.5058	0.2722	0.3239	0.0477	0.7970	**1.80**
2	**78.77**	**4.25**	0.1305	0.5009	**0.2752**	**0.3281**	**0.0489**	0.8154	1.82
4	78.74	4.44	**0.1231**	0.4881	0.2480	0.3161	0.0480	0.8599	1.95
8	77.88	4.31	0.1272	**0.4823**	0.2450	0.3195	0.0459	**0.8730**	2.01

individual behavior recognition in capturing relationships among traffic participants and demonstrate the necessity of independently evaluating multi-agent interaction understanding.

By integrating complementary RGB-Event information, our framework reduces P→B Error to 23.06% and B→I Error to 29.35%, corresponding to reductions of 1.04 and 3.98 percentage points over RGB-only Qwen3-VL-4B, respectively. The larger reduction in B→I Error indicates that our framework exhibits fewer interaction understanding failures among instances where the associated path actions are correctly recognized. This suggests that combining appearance semantics with fine-grained dynamic cues can benefit the understanding of relationships between interacting agents beyond individual behavior recognition. Overall, these results demonstrate the diagnostic value of TRACE-Bench in exposing cross-level understanding failures.

A.2.2 Performance Under Different Illumination Conditions.

We compare the fine-tuned models under daytime and nighttime conditions to investigate the contribution of event data across different illumination settings, as shown in Tab. 8 and 9. During daytime, RGB-only Qwen3-VL-4B outperforms its Event-only counterpart in accuracy (71.59 vs. 68.93) and BERT-F1 (0.3016 vs. 0.2791), reflecting the importance of dense appearance and scene semantics for traffic understanding. At night, however, the accuracy gap narrows from 2.66 to 1.41 percentage points, and Event-only surpasses RGB-only across all four open-ended QA metrics. This shift is consistent with the complementary value of event sensing under challenging illumination, where its sensitivity to intensity changes and high dynamic range can provide motion-related cues that may be less reliable captured by RGB frames. Moreover, Event-only achieves lower T-MAE under both conditions (5.15 vs. 5.30 during daytime and 3.80 vs. 3.87 at night), suggesting that its fine-grained temporal cues may benefit the localization of behavior and interaction intervals. By integrating both modalities, our method achieves the highest accuracy and lowest T-MA and spatiotemporal grounding error under both conditions. Compared with RGB-only, it improves accuracy by 0.37 and 0.80 percentage points during daytime and nighttime, respectively, supporting the benefit of combining RGB semantics with complementary event dynamics for fine-grained traffic understanding.

Table 11: Ablation of temporal and spatial pruning ratios.

Temporal	Spatial	Acc. ↑	T-MAE ↓	S-L1 ↓	ST Err. ↓	BERT-F1 ↑	ROUGE-L ↑	BLEU-4 ↑	CIDEr ↑	Latency ↓
25%	50%	78.65	4.35	0.1346	0.5176	0.2701	0.3243	0.0478	0.795	1.95
50%	50%	**78.77**	4.25	**0.1305**	0.5009	0.2752	0.3281	**0.0489**	0.8154	1.82
75%	50%	78.60	4.19	0.1342	**0.4918**	0.2598	0.3257	0.0458	0.8338	1.61
50%	25%	78.11	4.19	0.1336	0.5007	0.2732	**0.3300**	0.0466	**0.8476**	1.85
50%	75%	78.62	4.21	0.1357	0.5102	**0.2864**	0.3266	0.0473	0.7079	1.69
50%	90%	77.29	**4.14**	0.1331	0.5162	0.2547	0.3201	0.0461	0.7291	**1.54**

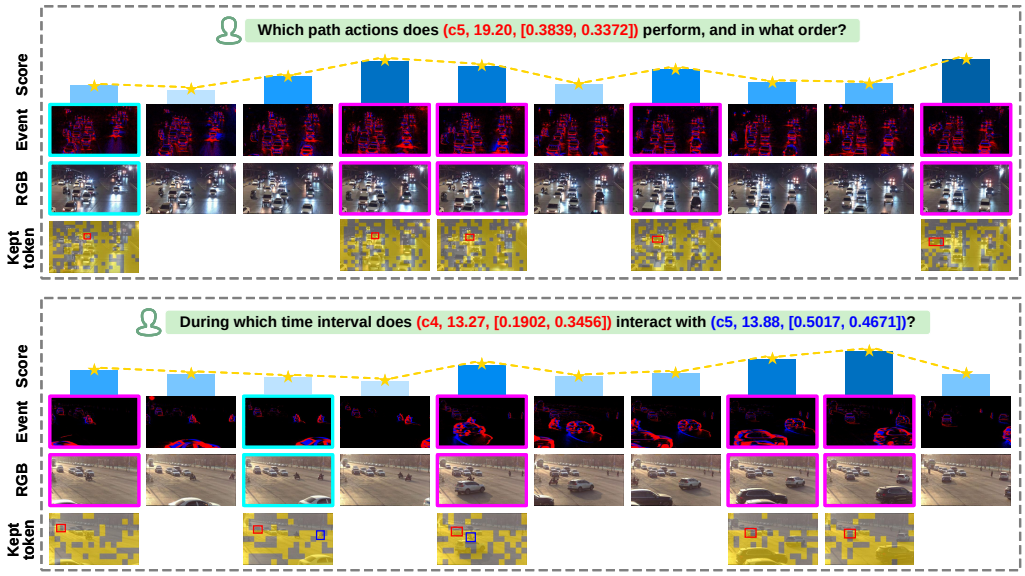


Figure 9: Visualization of spatiotemporal token selection in nighttime (top) and daytime (bottom) roadside scenes. The bars indicate event saliency scores for candidate frames. Purple and cyan borders denote frames retained by event-guided selection and uniform sampling, respectively. Retained spatial tokens are highlighted in yellow, with bounding boxes indicating the queried agents.

A.2.3 Qualitative Analysis of Interaction Understanding

Fig. 7 and 8 present qualitative comparisons of interaction-process descriptions under daytime and nighttime conditions. The RGB-only Qwen3-VL-4B captures partial spatial relationships but provides limited information about the behaviors underlying the interactions. In the daytime example, it merely identifies the relative positions of two vehicles, overlooking the cut-in maneuver and subsequent deceleration. In the nighttime example, it describes the relative lane positions of the two agents without identifying the slower leading vehicle as the motivation for overtaking. In contrast, our method more closely aligns with the annotated interaction processes by describing both the initiating behaviors and the affected agents' responses. These examples illustrate that our framework supports a more detailed understanding of traffic interactions beyond static spatial relationships.

A.2.4 More Ablation Study

Input Frame Sampling Rate. As shown in Tab. 10, the text-only setting achieves only 33.17% accuracy, substantially below all visual-input settings, highlighting the importance of visual evidence for traffic understanding. Increasing the sampling rate from 1 to 2 FPS improves performance, suggesting that the additional temporal observations help capture fine-grained spatiotemporal cues. However, further increasing FPS yields different effects across tasks. Denser sampling generally benefits spatial and spatiotemporal grounding, with 4 FPS achieving the lowest S-L1 (0.1231) and 8 FPS achieving the lowest ST Err. (0.4823), potentially due to finer observation of agent locations and their temporal changes. In contrast, accuracy and BERT-F1 do not improve monotonically, while T-MAE fluctuates across sampling rates. This suggests that fine-grained reasoning depends not

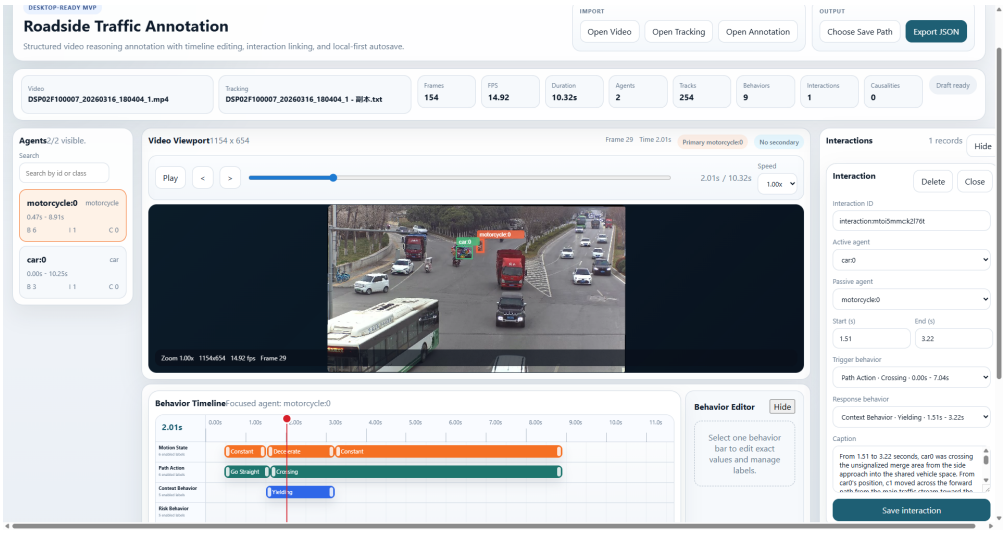


Figure 10: Annotation interface developed for TRACE-Bench. The tool jointly presents the source video, persistent agent records, temporal behavior tracks, and directed interaction editor. Annotators can edit motion states, path actions, context behaviors, interaction roles, time intervals, and descriptions within one synchronized workspace.

only on temporal density but also on leveraging informative observations. Overall, 2 FPS achieves the highest accuracy (78.77%), lowest T-MAE (4.25), and highest BERT-F1 (0.2752), providing a balanced setting for the diverse reasoning tasks in TRACE-Bench.

Spatiotemporal Pruning Ratios. As shown in Tab. 11, temporal and spatial pruning affect different aspects of traffic understanding. Increasing temporal pruning from 25% to 75% improves temporal and spatiotemporal grounding, but overly aggressive pruning reduces accuracy and BERT-F1, suggesting that fewer observations may preserve salient dynamics while losing contextual information needed for semantic reasoning. Spatial pruning exhibits a different trade-off: moderate pruning maintains competitive grounding and QA performance, whereas removing 90% of spatial tokens reduces accuracy and open-ended QA quality, potentially discarding task-relevant visual evidence. Overall, the 50%/50% configuration achieves the highest accuracy (78.77), lowest S-L1 (0.1305), and highest BLEU-4 (0.0489), providing a balanced setting across the diverse tasks.

A.2.5 Visualization of Spatiotemporal Selection

Fig. 9 visualizes the proposed spatiotemporal selection under nighttime and daytime conditions. Event-guided selection prioritizes frames with salient dynamic responses, while uniform sampling provides complementary observations to maintain temporal coverage. At the spatial level, the retained tokens include regions around the queried agents and their surrounding traffic context, while reducing tokens from less relevant areas. These examples illustrate how the two-stage selection combines dynamic saliency, temporal coverage, and question relevance to retain visual evidence for behavior and interaction understanding.

A.3 Dataset

A.3.1 Annotation Tool

To support fine-grained and temporally consistent annotation, we developed a dedicated tool for traffic behavior and interaction annotation, applicable to both roadside and driving scenarios, as shown in Fig. 10. The tool takes video sequences and corresponding object trajectories as input, allowing annotators to track persistent agent identities and frame-wise locations throughout each sequence. For behavior annotation, annotators assign behavior labels and specify their temporal intervals. For interaction annotation, the interface links initiators and affected agents, records response behaviors and interaction intervals, and supports natural-language descriptions of the interaction process. The

Table 12: Overview of the TRACE-Bench task hierarchy. *MC*: multiple choice; *Judgment*: yes/no question; *Loc.*: Spatiotemporal localization; *Temp.*: temporal localization.

Task Category	Description	Question Type
L1: Static Scene Understanding		
Road Type Recognition	Recognizes the visible main-road type.	MC
Motor-Vehicle Lane Count	Counts motor-vehicle lanes in the main roadway.	MC
Traffic Guidance Recognition	Recognize traffic signs.	MC
Road Topology Description	Describes road layout and traffic guidance.	Caption
L2: Agent Perception		
Category Recognition	Recognizes the category of a specified agent.	MC
Spatiotemporal Grounding	Localizes an agent at its first and last visible frames.	Loc.
L3: Behavior Understanding		
Behavior Recognition	Recognizes motion, path, or contextual behavior.	MC
Temporal Grounding	Localizes a behavior interval.	Temp.
Behavior Chain Understanding	Identifies the ordered path actions of an agent.	MC
L4: Directed Interaction Reasoning		
Interacting-Agent Identification	Determine whether an interaction occurs.	Judgment / MC
Temporal Interaction Grounding	Localizes an interaction interval.	Temp.
Response Behavior Understanding	Identifies the affected agent's response.	MC
Interaction Location Description	Describes where the interaction occurs.	Caption
Interaction Process Explanation	Describes the influence–response process.	Caption / MC
L5: Counterfactual and Hallucination Evaluation		
Counterfactual Reasoning	Infers the plausible outcome of hypothetical change.	MC
Hallucination Evaluation	Rejects evidence-inconsistent premises.	Judgment / MC

structured interface facilitates consistent cross-level associations among agents, behaviors, and interactions, and exports all annotations in a unified format for subsequent consistency checks and quality assurance.

A.3.2 QA Construction

For the first four tasks (static-scene understanding, agent perception, behavior understanding and directed interaction reasoning), manually designed templates are populated from the corresponding structured fields, preserving a direct evidence path from each answer to the annotation. Counterfactual questions modify a local condition of an observed behavior or interaction while preserving the remainder of the scene context. Hallucination samples introduce controlled inconsistencies, such as nonexistent agents, behaviors, or interactions, reversed influence directions, and invalid temporal conditions. For multiple-choice questions, GPT-5.5 generates or selects distractors that are semantically similar to the correct answer but inconsistent with the structured evidence, reducing reliance on superficial linguistic cues. It is also used to improve QA diversity and consistency while preserving the original facts and answer semantics.

A.3.3 Quality Control

We employ a multi-stage quality-control procedure throughout structured annotation and QA construction. To reduce ambiguity in behavior and interaction annotation, annotators follow unified guidelines that specify the definitions and observable criteria for each behavior category and directed interaction relation. Interactions are identified based on the initiating behavior and the observable response of the affected agent, rather than agent co-occurrence alone. Ambiguous behavior categories and interaction relations are further reviewed and resolved through expert discussion based on visual evidence. To improve temporal annotation consistency, we establish explicit criteria for behavior boundaries. The start time is defined as the earliest point at which sufficient observable evidence supports the behavior label, while the end time corresponds to the point at which the behavior ceases or transitions into another identifiable state. Following these criteria, three annotators independently annotate the start and end times of each behavior. Their annotations are then averaged to obtain the final temporal boundaries, reducing dependence on any single annotator's subjective

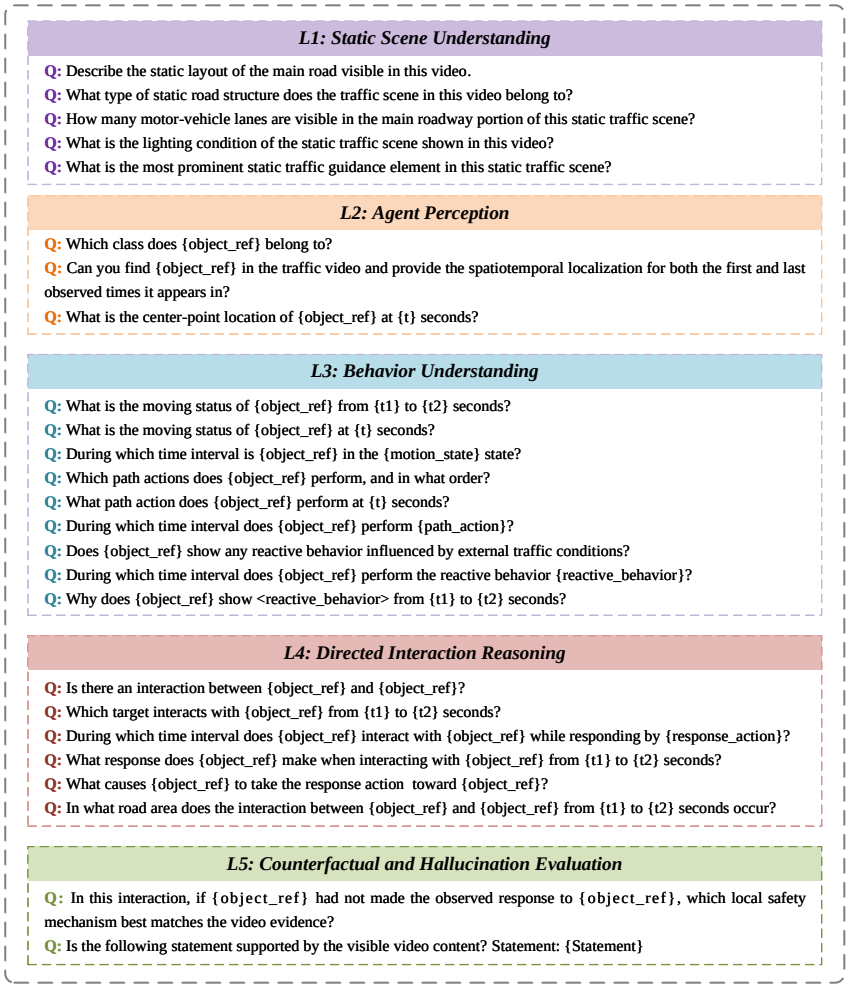


Figure 11: QA templates for the five TRACE-Bench task levels.

judgment. In addition, all generated QA pairs are further cross-checked against their corresponding structured evidence, with particular attention to question premises, answer uniqueness, temporal consistency, and visual answerability. This process helps reduce samples that lack sufficient visual evidence, contain semantic ambiguity, or include incorrectly labeled premises.

A.3.4 Detailed Task Definitions

TRACE-Bench evaluates traffic understanding along a progressive pathway from basic scene and agent perception to behavior understanding and interaction reasoning. Tab. 12 summarizes the definitions of all tasks.

A.3.5 QA Templates

Fig. 11 presents the representative template library used to cover the five TRACE-Bench task levels. Templates instantiate questions from structured annotations.

A.3.6 Prompt Templates

For reproducibility, we provide the exact multiple-choice distractor construction prompt, and QA language polishing prompt, as shown in Fig. 12 and 13. They support controlled language processing and option refinement.

Multiple-Choice Distractor Construction:

You are constructing multiple-choice QA for roadside traffic video understanding.
Given a question, its ground-truth answer, structured evidence, target references, generate exactly three incorrect but plausible distractors.

Inputs:
- Task type: {task_type}
- Question: {question}
- Ground-truth answer: {correct_answer}
- Target reference(s): {target_references}
- Evidence available to the annotator: {evidence}

Requirements:
1. Produce exactly three distractors. Together with the ground-truth answer, the four options must be unique.
2. Each distractor must be grammatically complete and match the answer format of the correct answer.
3. Keep the same local traffic context, target type, and abstraction level as the correct answer.
4. Make distractors plausible enough that they cannot be rejected using superficial wording alone.
5. Each distractor must contradict the visible evidence in at least one controlled aspect:
   - semantic class, motion state, path action, or response behavior;
   - interacting target or active/affected role;
   - temporal interval;
   - road area;
   - traffic-safety mechanism.
6. Do not introduce objects, actions, interactions, road areas, or causal relations unsupported by the available evidence, unless the task explicitly evaluates hallucination.
7. Do not use trivial negations, mutually exclusive wording cues, duplicate meanings, or vague alternatives such as “no risk”, “no need to adjust”, “larger gap”, “no path overlap”, or “outside the traffic stream”.
8. For path-action-chain questions, create distractors by reversing, swapping, or replacing actions while preserving a plausible sequence structure.
9. For interacting-target questions, use visible non-target agents as distractors.
10. For counterfactual questions, preserve the referenced targets and local road context, while changing the conflict mechanism, affected area, timing, object, or role.
11. Do not alter timestamps, normalized coordinates, target identifiers, or target-reference definitions supplied in the input.

Return JSON only:

```
{
 "distractors": [
   "<distractor 1>",
   "<distractor 2>",
   "<distractor 3>"
 ],
 "controlled_error_axes": [
   "<axis for distractor 1>",
   "<axis for distractor 2>",
   "<axis for distractor 3>"
 ]
}
```

Figure 12: The prompt template used for multiple-choice distractor construction.

QA Linguistic Polishing:

You are a quality-control editor for a roadside traffic video QA dataset.

Polish the QA item into concise, fluent English without changing its verified meaning.

Rules:
1. Do not add or remove evidence-supported facts, objects, actions, causal relations, or road features.
2. Preserve target identifiers, timestamps, coordinates, temporal boundaries, answer labels, and the correct option exactly.
3. Keep the question and answer semantically aligned and preserve the original answer type.
4. Keep four unique, equally plausible options for multiple-choice questions.

Input:
- Task type: {task_type}
- Question type: {question_type}
- Question: {question}
- Answer: {answer}
- Options: {options}
- Evidence: {evidence}

Return JSON only:

```
{
  "question": "<polished question>",
  "answer": "<polished answer>",
  "options": {"A": "...", "B": "...", "C": "...", "D": "..."}
}
```

Figure 13: The prompt template used for QA language polishing.