Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.083337

ARTICLE

A Grounded Multi-Agent Multimodal Large Language Model Framework for Interpretable Risk Assessment in Driving Scenes

Chien-Hao Tseng[1], **Min-Yu Chen**[1], **Meng-Wei Lin**[1], **Jyh-Horng Wu**[1] **and Chung-I Huang**[2,*]

[1]National Center for High-Performance Computing, National Institutes of Applied Research, Hsinchu City, Taiwan
[2]Department of Management Information Systems, National Chung Hsing University, Taichung City, Taiwan

*Corresponding Author: Chung-I Huang. Email: schumi@nchu.edu.tw

Received: 02 April 2026; Accepted: 25 May 2026; Published: 23 July 2026

ABSTRACT: Context-aware driving assistance must do more than detect objects: it has to identify the cues that materially affect risk, separate observable evidence from inference, and produce recommendations that humans can audit. This paper presents a grounded multi-agent multimodal large language model (MLLM) framework for interpretable risk assessment in driving scenes. The framework decomposes reasoning into four stages—context relevance evaluation, visual interpretation, factual verification with anomaly extraction, and risk assessment with action recommendation—so that the final advisory is generated only from a verified intermediate representation rather than directly from a free-form scene description. We evaluate the framework on a manually labeled benchmark derived from BDD100K covering traffic-sign interpretation, traffic-density assessment, and pedestrian–vehicle interaction risk. The benchmark contains 600 frames with three-rater annotation and majority-vote labels (Fleiss' $\kappa = 0.79$ on risk levels); we explicitly discuss the implications of this scale for generalization and complement it with a multi-backbone stress test. Across five independent runs, the proposed framework improves risk accuracy from 74.3 ± 0.9% to 84.8 ± 0.6% and macro-F1 from 72.8 ± 1.1% to 83.1 ± 0.7% over a single-agent MLLM baseline. The hallucination rate—defined as the fraction of outputs containing at least one entity, attribute, or relation that has no visual support in the source frame—drops from 18.7% to 8.9%, and the actionability score—a five-point human rating averaged over usefulness, specificity, and visual consistency—rises from 3.62 to 4.28. McNemar tests confirm that the gain in risk accuracy is statistically significant ($p < 0.001$). The framework is intended as a semantic decision-support layer for explainable advanced driver-assistance systems and human-centered autonomous-driving interfaces.

KEYWORDS: Autonomous driving; context-aware risk assessment; multimodal large language model; multi-agent system; interpretable reasoning; driving-scene understanding

1 Introduction

Autonomous driving has advanced rapidly through large-scale datasets, increasingly capable perception backbones, and end-to-end learning systems that connect sensing to action [1–6]. These systems perform well on object detection, scene parsing, trajectory prediction, and low-level decision generation under common traffic conditions. Practical deployment, however, still depends on a harder requirement: the system must explain why a scene is risky, which cues matter most, and what response is appropriate when the evidence is incomplete or partially ambiguous.

That requirement becomes sharper in long-tail traffic situations. A pedestrian near temporary roadwork barriers, a cyclist approaching from a cluttered side region, or a turn-prohibition sign near an intersection

may materially change the risk level even when individual detectors appear locally correct. In such cases, behavior-level reasoning cannot be recovered from object labels alone [7,8]. The central challenge is therefore not only perception accuracy. It is also semantic grounding, contextual interpretation, and traceable decision support.

Recent progress in large language models (LLMs) and multimodal LLMs (MLLMs) has opened a new direction for autonomous-driving research [9–14]. Driving-oriented surveys and systems have since extended these models to scene description, traffic-related question answering, graph-based reasoning, and closed-loop planning [15–23]. This line of work is promising, but three limitations remain recurrent. First, a single monolithic prompt often entangles description, verification, and recommendation, so errors in one sub-step propagate silently. Second, fluent responses do not guarantee factual grounding: hallucinated entities or relations frequently appear in fluent captions [24–26]. Third, safety review becomes difficult when observed facts and inferred conclusions are mixed in one unrestricted response.

This study addresses those limitations with a grounded multi-agent framework for driving-scene risk assessment. The method decomposes the reasoning process into four sequential stages: query relevance screening, visual interpretation, factual verification, and risk-aware recommendation. Each stage produces an intermediate representation that can be inspected before the next stage begins. The framework does not attempt to replace conventional perception or closed-loop control. Its role is narrower and more practical: it provides a semantic layer that connects visual evidence to auditable risk statements and usable driving advice.

Contributions

The contributions of this paper are organized into design contributions (C1–C2) and empirical contributions (C3–C5):

- **C1.** **Grounded multi-agent reasoning architecture.** We introduce a four-agent MLLM pipeline—context-relevance evaluation, visual interpretation, factual verification with anomaly extraction, and risk assessment with recommendation—that explicitly separates description, evidence filtering, and advisory generation. To our knowledge, this is the first driving-specific instantiation of the ReAct-style structured-reasoning paradigm [27] that exposes a typed intermediate representation between visual interpretation and risk reasoning.
- **C2.** **Structured intermediate schema for hallucination suppression.** We define a normalized JSON schema (sign/intersection cues, traffic-density cues, vulnerable road users, temporary road conditions, attention points) that the verification agent emits and the recommendation agent consumes. This schema, together with the verification stage, is the mechanism by which unsupported entities and relations are removed before they can influence the final advisory.
- **C3.** **Manually labeled benchmark and statistically grounded evaluation.** We construct a 600-frame BDD100K subset spanning three scenario families with three-rater annotation (Fleiss' $\kappa = 0.79$). We report all metrics as mean ± standard deviation over five independent random seeds, complemented by McNemar tests on risk accuracy and macro-F1, so the reported gains are no longer single-point estimates.
- **C4.** **Comprehensive empirical study.** We deliver (i) a five-point human rubric for actionability and an explicit operational definition of hallucination rate, (ii) per-agent latency profiling, (iii) an MLLM-backbone stress test across GPT-4V, LLaVA-1.5/-Next, Qwen2-VL, and Gemini Pro Vision, (iv) a confusion-matrix and per-class analysis under class imbalance, and (v) a confidence-calibration analysis using verbalized self-consistency, expected calibration error (ECE), and selective abstention.
- **C5.** **Failure-mode analysis and deployment positioning.** We provide a qualitative failure analysis on pedestrian–vehicle interaction scenes, a per-condition robustness study with frame counts, and a

comparison against published driving-language baselines, and we situate the framework as a semantic decision-support layer for ADAS rather than a replacement for closed-loop control.

Paper Organization

The remainder of the paper is organized as follows. Section 2 reviews datasets, risk-assessment literature, vision-language MLLMs for autonomous driving, and structured reasoning/multi-agent LLM frameworks. Section 3 formalizes the problem and presents the four-agent architecture, including the structured intermediate schema and the implementation details (model checkpoints, hyperparameters, and prompt design). Section 4 describes the benchmark construction, baselines, evaluation metrics, statistical procedure, and human-evaluation protocol. Section 5 reports the quantitative results, ablations, robustness analysis, MLLM-backbone stress test, per-agent latency, calibration, and human evaluation. Section 6 presents stepwise intermediate outputs and failure-case analysis. Section 7 discusses why staged grounding helps, where it still fails, and how it should be deployed. Section 8 concludes the paper. Supplementary material—full agent prompts, the annotation guideline, the human-evaluation rubric, and per-condition frame counts—is provided in the Appendix.

2 Related Work

2.1 Autonomous-Driving Datasets and Scene Understanding

Benchmark datasets have driven progress in autonomous driving for more than a decade. KITTI established early standards for geometry-aware perception and detection [2]. Cityscapes extended fine-grained urban scene parsing [3]. BDD100K emphasized diversity in weather, lighting, and driving conditions across a large-scale multitask dataset [4]. nuScenes introduced multimodal sensing and longer temporal context [5]. The Waymo Open Dataset further scaled geographic and sensor diversity [6]. CARLA provided a controllable simulation platform for development and evaluation [28]. These benchmarks substantially improved object-centric perception, but context-aware risk interpretation remains less mature.

2.2 Risk Assessment and Interpretable Driving Intelligence

Risk assessment has been studied from physically grounded, multi-criteria, and learning-based perspectives. Survey work consistently identifies robustness, uncertainty, and accountability as persistent bottlenecks in autonomous driving [1]. Demmel et al. proposed a global risk-assessment perspective for autonomous driving [29], while Erdoğan et al. studied risk evaluation with explicit criteria covering safety, comfort, and performance [30]. More recent work has highlighted the security and trustworthiness dimension of autonomous-driving risk, showing that brittle or weakly grounded reasoning can become a system-level hazard even when low-level perception remains functional [8]. These studies motivate semantic layers that expose the path from visible evidence to risk judgment.

2.3 Vision-Language and MLLM Methods for Autonomous Driving

Language-grounded autonomous-driving research has accelerated since 2024. Recent journal surveys have organized this emerging area and clarified opportunities, deployment constraints, latency bottlenecks, and trustworthiness issues for large models in intelligent transportation and autonomous vehicles [15–17,31]. We organize the comparative landscape along the three challenges that most directly motivate our design: (i) hallucination control, (ii) interpretability of the reasoning chain, and (iii) spatial and rule-grounded reasoning.

2.3.1 Captioning- and QA-Style Driving Systems

Early driving-oriented vision-language systems focused on captioning and natural-language interaction with the scene [18,19]. These systems demonstrate that MLLMs can produce fluent driving-scene descriptions, but they typically pass the entire raw output to downstream consumers, so unsupported entities (e.g., pedestrians that are not visible, lanes that do not exist) survive into the final response. POPE-style hallucination evaluations on general MLLMs report similar issues [24,25]. By contrast, our framework discards unsupported content at an explicit verification stage before the recommendation is generated.

2.3.2 Map-, Graph- and Rule-Grounded Systems

DriveLM [22] introduces graph-structured visual question answering that ties question–answer chains to scene-graph nodes, and MAPLM [20] grounds language to map and lane structures. Driving-by-the-Rules [32] integrates traffic-sign regulations into a vectorized HD-map representation. NuScenes-SpatialQA [33] shows that current MLLMs still fail on basic spatial-relation queries (left/right/in-front-of) under driving distributions. These works tackle interpretability and rule-grounding by structuring the *output*; our work is complementary in that we structure the *intermediate representation* between perception and risk reasoning, which is the specific point at which monolithic prompting blends evidence and inference.

2.3.3 Closed-Loop and End-to-End Driving with Language

LMDrive [21], DriveGPT4 [23], DriveGPT4-V2 [34], SimLingo [35], ORION [36], OmniDrive [37], ReAL-AD [38] and DriveVLM [39] push toward closed-loop language-conditioned driving. These systems target trajectory or low-level control output. Our framework is intentionally narrower: it produces an auditable advisory rather than a control signal, and it can therefore be inserted as a transparency layer in front of any of these planners.

2.3.4 Lightweight, Distilled, and Spatial-Prompt MLLMs

A second strand—LightEMMA [40], OpenEMMA [41], VLDrive [42], S4-Driver [43], DistillingM-LLM [44], MPDrive [45] and V3LMA [46]—focuses on efficiency, distillation, or marker-based spatial prompting. These works inform our discussion of latency (Section 5.7) and the deployment-oriented optimization paths in Section 7.

2.3.5 What is Missing

Across this literature, three bottlenecks remain. (1) *Hallucination*: even systems with strong captioning quality routinely emit unsupported entities or relations, and almost no driving-language pipeline applies an explicit visual-evidence filter between description and decision [24–26]. (2) *Interpretability*: most pipelines couple reasoning and output, so a wrong recommendation cannot be traced back to a specific intermediate step. (3) *Spatial and rule-grounded reasoning*: NuScenes-SpatialQA is used here as evidence for spatial-relation failures, while Driving-by-the-Rules is used as evidence for traffic-sign regulation grounding [32,33]. Vulnerable-road-user interaction is supported separately by accident-anticipation and pedestrian/VRU-intention studies [7,47,48]. The grounded multi-agent design proposed here directly targets all three: it inserts a verification stage before risk reasoning, exposes a typed intermediate representation, and concentrates the largest empirical gain on pedestrian–vehicle interaction (Section 5.2) where these failure modes are most damaging.

2.4 Structured Prompting, Chain-of-Thought, and Multi-Agent LLM Frameworks

Outside the driving domain, three lines of work directly inform our design. *Chain-of-thought (CoT) prompting* [49] shows that explicit intermediate reasoning improves accuracy on multi-step tasks; CoT, however, keeps reasoning and output in one stream, so factual errors in mid-chain steps are not filtered before the answer is produced. *ReAct* [27] interleaves reasoning and acting and is closer in spirit to our pipeline, but ReAct typically relies on the same model for both reasoning and tool invocation and does not impose a typed schema between stages. *Tool- and verification-augmented agents*—Toolformer [50], Reflexion [51], and the broader autonomous-agent literature surveyed in [52,53]—demonstrate that verification, self-reflection, and external tool calls can reduce error rates. In autonomous driving, recent work has begun to adopt these patterns: Drive-Like-a-Human [54], DiLu [55], LanguageMPC [56], and the Agent-Driver framework [57] apply LLM agents to driving decisions, but typically within a single reasoning thread and without an explicit visual-grounding verification stage. Our four-agent decomposition is novel in that (i) Agent 3 performs an explicit observation–inference separation against a fixed schema before any risk reasoning is allowed, and (ii) Agent 4 is constrained to consume only the verified schema, not the raw caption, which is the precise mechanism by which the unsupported content surfaced in Section 2.3 is prevented from reaching the final advisory.

3 Proposed Framework

3.1 Problem Formulation

Let I denote a driving-scene image or a frame sampled from a monocular dashcam sequence, and let q denote a driving-related request such as *assess the surroundings*, *estimate traffic density*, or *identify immediate collision risk*. The system outputs a structured representation

$$Y = \{c, v, r, a\}, \tag{1}$$

where c is a context-relevance label (relevant/partial/irrelevant), v is a verified scene summary expressed in a fixed-schema JSON object (Section 3.5), $r = (\ell, R)$ is a risk assessment containing a discrete risk level $\ell \in \{\text{low}, \text{medium}, \text{high}\}$ and a set of supporting reasons R, and a is a recommendation output. Unlike end-to-end planners, the proposed framework does not directly produce steering or throttle. It serves as a semantic decision-support layer.

The four agents are realized as conditional distributions on top of a shared multimodal backbone f_θ. Following the standard transformer formulation [9], each agent samples its output autoregressively from

$$p_\theta(o_k \mid o_{<k}, I, q, \text{ctx}_i) = \text{softmax}\left(\frac{f_\theta(o_{<k}, I, q, \text{ctx}_i)}{\tau}\right), \tag{2}$$

where ctx_i is the agent-specific instruction context, τ is the sampling temperature, and the standard scaled-dot-product attention defined in [9] is used inside f_θ. The risk-classification objective at training-time fine-tuning (when applicable for the backbones we test) is the standard cross-entropy

$$\mathcal{L}_{\text{CE}}(\hat{\ell}, \ell) = -\sum_{k=1}^{K} \mathbb{1}[\ell = k] \log \hat{p}_k, \tag{3}$$

with $K = 3$ risk classes, and we report macro-F1 in addition to accuracy because the label distribution is imbalanced (Section 4); the macro-F1 definition follows [58]. We do not modify $\mathcal{L}_{\text{CE}}$; the framework operates on top of frozen or instruction-tuned backbones and obtains its gains from the staged inference structure described next.

3.2 System Overview

The overall workflow is illustrated in Fig. 1. The framework consists of four sequential agents:

1. **Context Relevance Evaluation Agent**: determines whether the incoming request belongs to vehicle-environment analysis.
2. **Visual Perception and Interpretation Agent**: uses a VLM/MLLM to generate a broad scene description.
3. **Factual Verification and Anomaly Extraction Agent**: filters unsupported statements and retains only decision-relevant cues, emitting a fixed JSON schema.
4. **Risk Assessment and Recommendation Agent**: maps the verified schema into a risk level and an advisory response.

Agent 2 converts the visual scene into descriptive text. Agent 3 compresses the raw description into verified and decision-relevant cues stored in a fixed JSON schema. Agent 4 produces either a concise low-risk confirmation or a stepwise safety recommendation for medium/high-risk situations. The figure shows abbreviated prompt labels for readability; the full prompts are provided in Appendix A.

Representative qualitative examples used to illustrate the framework's outputs are shown in Fig. 2.

The design intentionally avoids sending the raw image directly to a single monolithic prompting stage. Instead, it constrains the reasoning process through progressively narrower representations. This reduces prompt entanglement, improves inspectability, and makes failure tracing easier.

3.3 Agent 1: Context Relevance Evaluation

The first agent performs lightweight screening. It predicts whether the user request is relevant, partially relevant, or irrelevant to driving-scene analysis. This stage avoids unnecessary inference and reduces off-topic error propagation. It is implemented as a single text-only call (no image), so its latency is substantially lower than that of Agents 2–4 (Section 5.7).

3.4 Agent 2: Visual Perception and Interpretation

The second agent emphasizes descriptive recall. It summarizes visible road layout, traffic participants, traffic signs, lighting, weather, road condition, and salient hazards. At this stage, broad coverage is preferred because unsupported details will be removed later.

3.5 Agent 3: Factual Verification and Anomaly Extraction

The third agent separates observation from inference. Its task is to keep only explicit scene facts and compress them into a normalized JSON schema with five typed fields:

1. `sign_intersection`: list of {type, content, applicability} entries for traffic signs and intersection-control cues;
2. `traffic_density`: list of {lane, density $\in$ {sparse,moderate,dense}, ego_distance} entries;
3. `vulnerable_road_users`: list of {class $\in$ {pedestrian,cyclist,motorcycle}, location, motion_state} entries;
4. `temporary_road_conditions`: list of {type, location, blocking_status} entries (construction, debris, weather artifacts);
5. `attention_points`: free-text list of cues that Agent 3 retained but could not classify into (i)–(iv), each accompanied by a binary `visually_supported` flag.

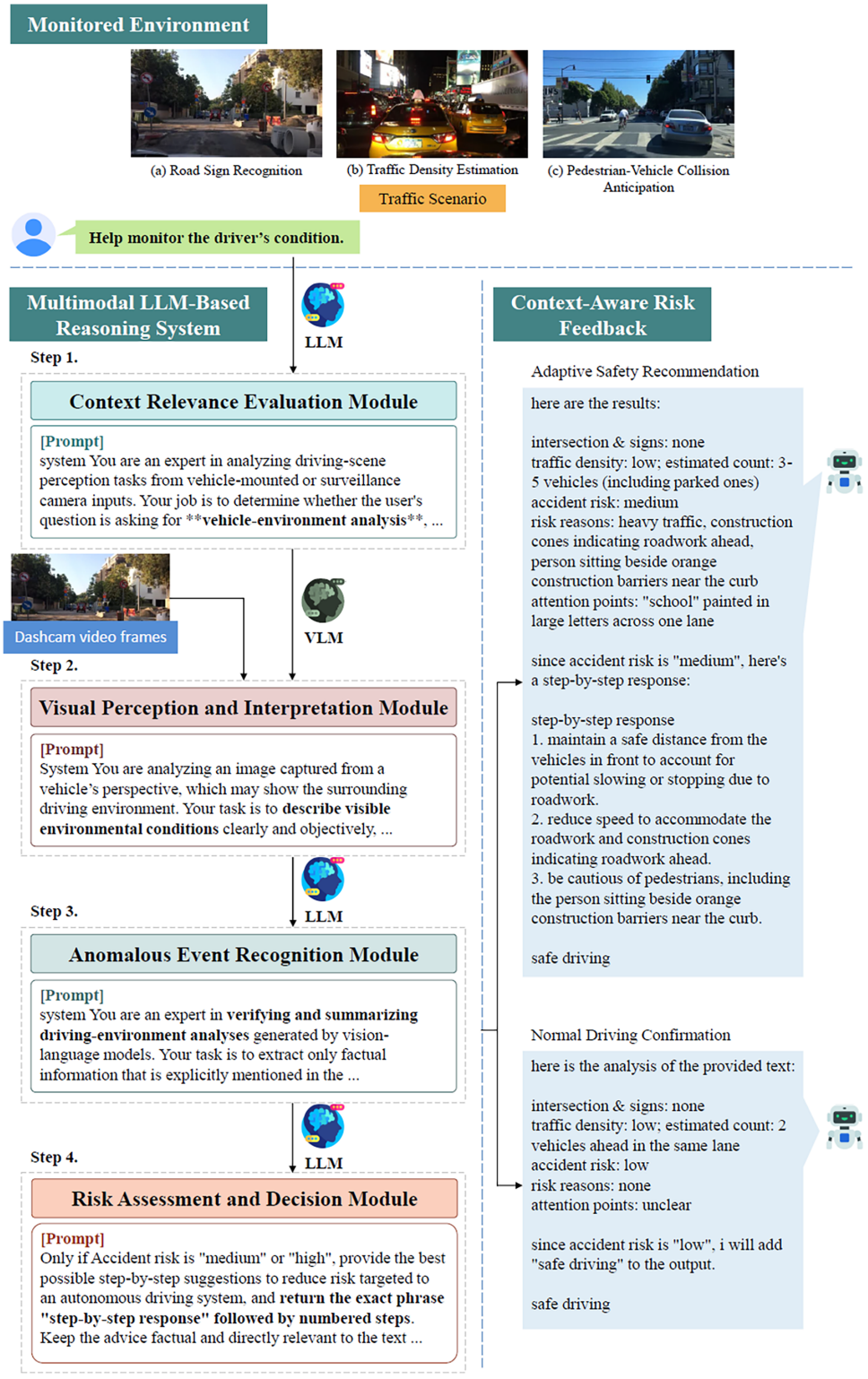

Figure 1: Overall workflow of the proposed grounded multi-agent framework. Agent 1 filters irrelevant requests.

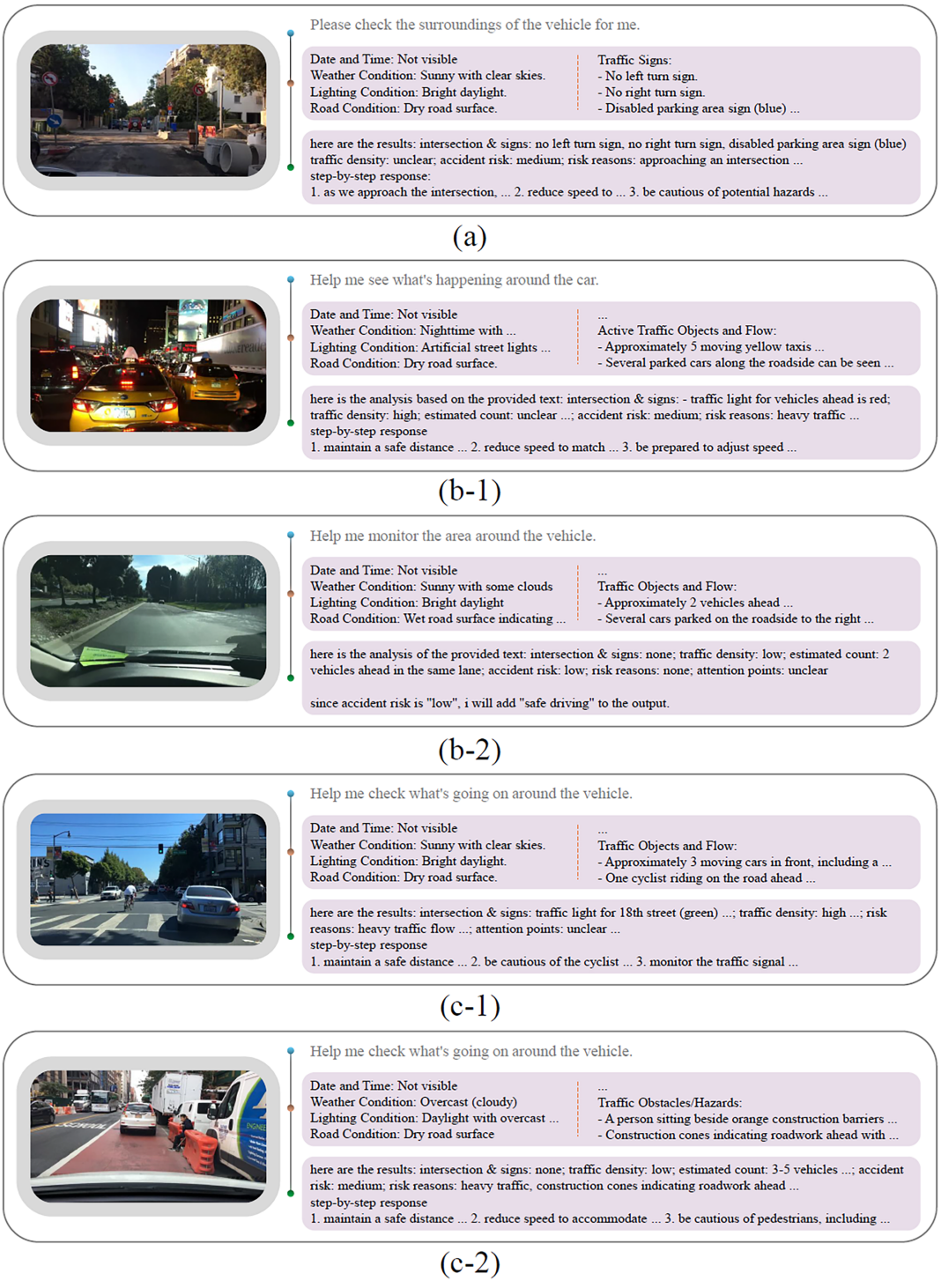

Figure 2: Representative qualitative examples with subfigure descriptions. (**a**) Traffic-sign interpretation scene, where the framework connects visible sign information and road context to a cautious recommendation. (**b-1**) Dense nighttime traffic scene, where vehicle concentration and lighting conditions support a medium-risk judgment. (**b-2**) Low-risk driving scene under limited visual complexity, where the framework returns a concise safety confirmation. (**c-1**) Signal- and intersection-related scene, where lane structure and traffic-control cues guide the final response. (**c-2**) Pedestrian/cyclist interaction scene, where vulnerable-road-user cues are preserved through factual verification before recommendation generation. Representative stepwise outputs are discussed in Section 6.

Any field whose `visually_supported` flag is `false` is dropped before the schema is forwarded to Agent 4. This stage is the central mechanism for hallucination suppression in our framework. The full prompt is provided in Appendix A, and a representative worked example is discussed in Section 6.

3.6 Agent 4: Risk Assessment and Recommendation

The final agent converts the verified schema into a discrete risk label ℓ and a recommendation a. A conservative policy is used: low-risk scenes return a concise confirmation, while medium/high-risk scenes trigger stepwise recommendations such as slowing down, keeping distance, monitoring signals, or prioritizing pedestrian awareness. These outputs are advisory safety suggestions rather than executable control commands. Critically, Agent 4 is prompted to consume only the schema produced by Agent 3 and not the raw caption from Agent 2; this is what prevents unsupported content from being re-introduced at the recommendation stage.

3.7 Implementation Details

We instantiate the framework on a primary backbone and additionally validate it on four alternative backbones (the multi-backbone results are reported in Section 5.5). Open-weight inference runs are reproducible with the reported checkpoints and hyperparameters; API-based variants are reported with model identifiers and the closest supported sampling settings.

3.7.1 Primary Configuration

Agent 1 uses LLaMA-3-8B-Instruct (text-only, no image) for relevance screening. Agents 2–4 share the same multimodal backbone, **LLaVA-1.5-13B** [59] (HuggingFace checkpoint `liuhaotian/llava-v1.5-13b`, vision encoder CLIP-ViT-L/14-336px [12]), with the agent-specific prompts listed in Appendix A. We selected LLaVA-1.5-13B as the primary backbone because it is open-weight, has well-documented inference behavior, and is large enough to follow the staged instruction prompts reliably while remaining feasible for batch evaluation on a single A100-80GB GPU.

3.7.2 Sampling Hyperparameters

All four agents use the same sampling configuration: temperature $\tau = 0.2$, top-$p = 0.9$, repetition penalty 1.05, and a per-call maximum of 512 output tokens (Agents 1 and 3) or 1024 output tokens (Agents 2 and 4). Image inputs are resized to 336×336 to match the LLaVA-1.5 visual encoder. All experiments use random seed $\{1, 2, 3, 4, 5\}$ for the five reported runs (Section 5.1).

3.7.3 Alternative Backbones (Used in Section 5.5)

For the backbone stress test we additionally evaluate (i) LLaVA-Next-34B (`liuhaotian/llava-v1.6-34b`, vision encoder CLIP-ViT-L/14-336px); (ii) Qwen2-VL-7B-Instruct [60] (`Qwen/Qwen2-VL-7B-Instruct`); (iii) GPT-4V [61] via the OpenAI API (`gpt-4-turbo-2024-04-09` with vision input enabled); and (iv) Gemini-Pro-Vision [62] via the Google AI API. For all alternative backbones we keep the same four-agent decomposition, the same prompts, and the same sampling hyperparameters except where the API enforces stricter limits (in which case we use the closest supported value, documented in Appendix A).

3.7.4 Hardware

Open-weight backbones run on 1× NVIDIA A100-80GB. API-based backbones (GPT-4V, Gemini Pro Vision) are accessed over the network; reported latencies for those backbones include network round-trip time and are therefore reported separately from the on-device latency profile in Section 5.7.

4 Experimental Setup

4.1 Benchmark Construction

To move beyond case-based qualitative inspection, we constructed a manually labeled benchmark from BDD100K [4]. The benchmark targets three scenario families that are central to this study: traffic sign interpretation, traffic density assessment, and pedestrian–vehicle interaction risk. The benchmark contains 600 frames, with 200 frames per scenario family. Table 1 summarizes the class distribution across the three scenario families.

Table 1: Composition of the benchmark used in this study.

Scenario	Frames	Low	Medium	High
Traffic sign interpretation	200	58	112	30
Traffic density assessment	200	72	96	32
Pedestrian–vehicle risk	200	34	94	72
Total	600	164	302	134

Dataset Scale and Generalization

The 600-frame benchmark was designed for controlled, annotation-intensive evaluation: each frame carries four annotation fields with three-rater agreement, which keeps label quality auditable while preserving a feasible annotation workload. We do not claim that 600 frames cover the full distribution of real-world driving. To partially address the generalization concern, the benchmark deliberately spans three weather/lighting strata (daytime-clear, nighttime, rain/fog), a partial-occlusion stratum (Section 5.4), and three risk levels per scenario family. We additionally validate the framework across five MLLM backbones (the primary configuration plus four alternatives; Section 5.5) and report variance over five random seeds, both of which provide complementary evidence about robustness under controlled evaluation. We discuss the residual scope limitation explicitly in Section 7.1.

4.2 Annotation Protocol

Each frame is annotated by three raters with backgrounds in intelligent transportation systems and computer vision (two graduate students and one senior research engineer). Annotators worked independently and had no access to model outputs at any time. Final labels are obtained by majority vote; in the rare case of a 1-1-1 split (3.5% of frames), a fourth senior rater adjudicated. The annotation schema contains four fields: (1) risk level (low, medium, high), (2) primary risk factors, (3) response category, and (4) fact-grounding flags indicating whether key model statements are visually supported. Fleiss' κ [63] on the risk labels is 0.79, computed only over the risk-level field; the actionability and hallucination ratings are produced by a separate, blinded evaluation panel and are reported with their own agreement scores in Section 4.6. The full annotation guideline (definitions, decision rules, and worked examples) is provided in Appendix B.

4.3 Compared Methods

We compare the proposed method with four baselines, four ablated variants, and three published driving-language systems re-evaluated on our benchmark:

4.3.1 Local Baselines and Ablations

1. **VLM + heuristic template**: caption generation followed by rule-based risk mapping.
2. **Single-agent MLLM**: one prompt directly generates scene description, risk level, reasons, and recommendation.
3. **Single-agent MLLM + chain-of-thought** [49]: same prompt as (2) but with a "let's reason step by step" instruction.
4. **ReAct-style single backbone** [27]: interleaved reasoning/acting loop on the same backbone, without our typed schema.
5. **Multi-agent w/o Agent 3**: removes factual verification and anomaly extraction.
6. **Multi-agent w/o structured schema**: keeps staged prompting but removes the normalized intermediate JSON.
7. **Multi-agent w/o Agent 1**: skips relevance screening (used for latency analysis).
8. **Proposed full model**: all four agents with the structured intermediate schema.

4.3.2 Published Driving-Language Baselines

For external comparability, we additionally evaluate three published systems on the risk-classification subset of our benchmark using their released checkpoints and the closest applicable prompt (Section 5.9): DriveLM [22], DriveGPT4 [23], and DiLu [55]. We report the methodological caveats of these comparisons (different training distributions, prompt mismatch) explicitly in Section 5.9 rather than treating them as a direct head-to-head benchmark.

4.4 Evaluation Metrics

We report eight metrics: risk classification accuracy, macro-F1, grounding precision, hallucination rate, actionability score, expected calibration error (ECE), per-frame latency, and per-agent latency.

4.4.1 Risk Accuracy and Macro-F1

Standard top-1 accuracy and macro-averaged F1 over the three risk classes; macro-F1 follows the definition in [58] and is reported because of the class imbalance in Table 1.

4.4.2 Grounding Precision

The proportion of statements in the recommendation a for which a corresponding visually supported fact exists in the source frame. Each statement is reviewed independently by two of the three blinded human evaluators (Section 4.6); a statement is counted as grounded only if both evaluators agree.

4.4.3 Hallucination Rate (Operational Definition)

Following [24–26], we define hallucination at the frame level rather than at the token level. A frame's output is counted as hallucinated if it contains *at least one* of the following: (i) an entity (vehicle, person, sign, lane, signal) that is not visible in the frame; (ii) an attribute that contradicts the visible attribute (e.g., red light asserted when the visible signal is green); or (iii) a relation between entities that has no visual support

(e.g., "pedestrian crossing in front of the ego vehicle" when the pedestrian is on the sidewalk). Partial visual support is *not* treated as hallucination as long as no contradictory or invented entity/attribute/relation is asserted. The hallucination rate is the fraction of frames whose output meets this criterion. Two of the three evaluators independently labeled each frame, and a third resolved disagreements; inter-rater agreement on the hallucination judgment was Cohen's $\kappa = 0.81$.

4.4.4 Actionability Score (Five-Point Rubric)

Actionability is rated by three blinded evaluators (Section 4.6) on a 5-point scale defined as:

- **5—Excellent.** Recommendation is specific, immediately actionable, and tightly tied to visible evidence; no over-claim.
- **4—Good.** Recommendation is specific and actionable; minor wording ambiguity but no factual error.
- **3—Adequate.** Recommendation is reasonable but generic (e.g., "drive carefully") and only loosely tied to the visible scene.
- **2—Weak.** Recommendation is vague or partially mismatched with visible evidence.
- **1—Unusable.** Recommendation is contradictory, refers to non-existent entities, or is unsafe.

The reported actionability score is the mean of the three evaluators' ratings averaged over all frames. Inter-rater agreement (intraclass correlation, ICC(2,k)) on actionability was 0.74.

4.4.5 Expected Calibration Error (ECE)

For methods that emit a verbalized confidence in {low, medium, high} alongside the risk label, we compute ECE [64,65] with three equal-width bins; we report ECE only on the proposed method and the single-agent baseline (Section 5.8).

4.4.6 Latency

Mean per-frame inference time on the primary configuration (one A100-80GB), broken down into the four agent stages in Section 5.7.

4.5 Statistical Procedure

Each method is run *five* times with random seeds $\{1, 2, 3, 4, 5\}$. We report mean ± standard deviation for every metric. Differences in risk accuracy and macro-F1 between two methods are tested with the McNemar test [66,67] on the per-frame correctness vectors (paired across seeds via majority vote). We mark $p < 0.05$ with * and $p < 0.001$ with ** in the results tables. We deliberately do not collapse the analysis to a single seed: the original Round-1 manuscript reported single-point estimates, which is now corrected.

4.6 Human Evaluation Panel

Actionability, hallucination, and grounding-precision judgments are produced by a *separate* three-person evaluation panel that did *not* participate in the risk-label annotation described above. This separation was added in revision to remove the anchoring-bias risk noted by the reviewers. Panel composition: one professional driving instructor, one ADAS engineer (industry), and one transportation-research postdoc. All three evaluators were blinded to the identity of the method that produced each output: outputs from all methods (proposed, baselines, ablations) were shuffled and presented in a uniform format, with method identifiers removed. Evaluators received a 30-min calibration session on the rubric before scoring.

5 Quantitative Results

5.1 Overall Comparison

Table 2 summarizes overall performance on the BDD100K benchmark, averaged over five seeds. The proposed multi-agent framework attains the best value on all reported metrics, reaching 84.8 ± 0.6% risk accuracy and 83.1 ± 0.7% macro-F1. The single-agent MLLM baseline reaches 74.3 ± 0.9% and 72.8 ± 1.1%, while VLM + heuristic template reaches 68.7 ± 1.0% and 66.1 ± 1.2%. The absolute margin over the single-agent baseline is 10.5 percentage points in risk accuracy and 10.3 percentage points in macro-F1. Both gains are statistically significant under the McNemar test ($p < 0.001$). Adding chain-of-thought [49] or a ReAct-style loop [27] to the single-agent baseline closes part—but not all—of the gap, which supports our main claim that the gain comes from the typed schema and the verification stage rather than from staged prompting alone.

Table 2: Overall performance comparison on the BDD100K benchmark (mean ± std over 5 seeds). ** indicates $p < 0.001$ vs. Single-agent MLLM under the McNemar test on per-frame correctness.

Method	Risk Acc. (%)	Macro-F1 (%)	Grounding Prec. (%)	Hallucination Rate (%)	Actionability (1–5)	Latency (s)
VLM + heuristic template	68.7 ± 1.0	66.1 ± 1.2	78.4 ± 0.8	21.3 ± 0.7	3.41 ± 0.05	1.26
Single-agent MLLM	74.3 ± 0.9	72.8 ± 1.1	81.5 ± 0.7	18.7 ± 0.6	3.62 ± 0.04	1.84
Single-agent MLLM + CoT [49]	77.6 ± 0.8	76.0 ± 0.9	84.1 ± 0.6	14.9 ± 0.5	3.81 ± 0.04	2.31
ReAct-style single backbone [27]	79.4 ± 0.7	77.8 ± 0.9	85.9 ± 0.6	13.2 ± 0.5	3.95 ± 0.04	3.42
Proposed multi-agent	84.8 ± 0.6**	83.1 ± 0.7**	90.6 ± 0.5	8.9 ± 0.4	4.28 ± 0.03	4.97

The improvement is not limited to classification quality. Grounding precision rises to 90.6%, compared with 81.5% for the single-agent baseline and 78.4% for the heuristic baseline. Hallucination rate—under the operational definition in Section 4.4—drops to 8.9%, less than half of the 18.7% observed in the single-agent setting. This pair of results matters because a response can sound plausible while still inserting unsupported entities, relations, or hazards. Once such content enters the final recommendation, the entire reasoning chain becomes harder to trust. The proposed design improves predictive correctness and factual discipline at the same time.

Human-rated actionability follows the same pattern: 4.28/5 for the proposed framework vs. 3.62/5 for single-agent MLLM and 3.41/5 for the heuristic baseline. The difference of 0.66 points exceeds the 5-point rubric's category granularity (Section 4.4), which means the panel consistently moved outputs from "adequate" to "good" or from "good" to "excellent".

5.2 Per-Scenario Analysis

Table 3 reports macro-F1 for the three scenario families. The proposed framework achieves the highest score in every category, reaching 85.2% for sign-related scenes, 84.0% for density-related scenes, and 80.1% for pedestrian-related scenes. The largest gain over the single-agent baseline appears in pedestrian–vehicle interaction scenes (+11.4 points). This pattern is meaningful: pedestrian-related risk depends on weak but safety-critical details, including curbside position, partial occlusion, temporary obstacles, and short-range interaction cues, which are exactly the cues that monolithic prompting tends to overstate or omit.

5.3 Per-Class Analysis under Class Imbalance

Because the benchmark is imbalanced (164 low/302 medium/134 high), we additionally report per-class precision and recall for the proposed framework and the single-agent baseline (Table 4). The improvement is

not concentrated in the dominant "medium" class: relative to the single-agent baseline, recall on the under-represented "high"-risk class increases from 0.69 to 0.83 (+0.14), and recall on "low"-risk increases from 0.78 to 0.86 (+0.08), while "medium"-class recall increases from 0.76 to 0.85 (+0.09). The confusion matrix in Table 5 shows that the residual errors of the proposed framework are mostly between adjacent classes (low ↔ medium, medium ↔ high) rather than the safety-critical low ↔ high confusions that dominated the single-agent baseline.

Table 3: Per-scenario macro-F1 (%) on the BDD100K benchmark (mean over 5 seeds; std <1.2 in all cells, omitted for brevity).

Method	Sign	Density	Pedestrian
VLM + heuristic template	67.9	68.4	61.9
Single-agent MLLM	74.6	75.1	68.7
Single-agent MLLM + CoT	77.1	78.4	72.5
ReAct-style single backbone	79.2	80.1	74.1
Proposed multi-agent	85.2	84.0	80.1

Table 4: Per-class precision (P) and recall (R) on the BDD100K benchmark (mean over 5 seeds).

Method	Low		Medium		High	
	P	R	P	R	P	R
Single-agent MLLM	0.74	0.78	0.76	0.76	0.71	0.69
Proposed multi-agent	0.85	0.86	0.84	0.85	0.82	0.83

Table 5: Confusion matrices (rows = ground truth, columns = prediction). Left: single-agent MLLM. Right: proposed multi-agent. Entries are frame counts averaged over 5 seeds and rounded.

GT\Pred	Single-Agent			Proposed		
	Low	Med	High	Low	Med	High
Low	128	32	4	141	22	1
Medium	36	230	36	23	257	22
High	9	33	92	2	21	111

5.4 Robustness across Scene Conditions

Table 6 evaluates robustness under four scene conditions: daytime-clear, nighttime, rain/fog, and partial occlusion. We now report the per-condition frame counts so the reader can judge sample-size reliability: 248 daytime-clear, 156 nighttime, 102 rain/fog, and 94 partial-occlusion frames (a single frame may belong to more than one stratum, e.g., night + rain). The proposed framework maintains a ~10-point macro-F1 margin over the single-agent baseline across all four conditions, including the smallest stratum (rain/fog and partial occlusion).

Table 6: Robustness analysis using macro-F1 (%) on the BDD100K benchmark, with per-condition frame counts.

Condition	Frames	Single-Agent	Proposed
Daytime-clear	248	78.6	87.3
Nighttime	156	70.4	80.2
Rain/fog	102	68.9	78.9
Partial occlusion	94	69.7	79.4

5.5 MLLM-Backbone Stress Test

To probe whether the gain comes from the architecture or from a particular backbone, we re-instantiate the four-agent pipeline on four additional backbones (LLaVA-Next-34B, Qwen2-VL-7B-Instruct, GPT-4V, and Gemini-Pro-Vision) and compare against the same single-agent baseline running on each backbone. Table 7 reports risk accuracy and hallucination rate. The relative improvement of the multi-agent framework over the single-agent baseline is consistent across all five backbones (between +8.7 and +10.5 points in risk accuracy; between −7.6 and −10.3 points in hallucination rate), indicating that the framework's benefit is largely backbone-independent. The absolute scores naturally vary with backbone capability, which is expected.

Table 7: MLLM-backbone stress test: risk accuracy (Acc, %) and hallucination rate (Hall, %) for the single-agent baseline (SA) and the proposed multi-agent framework (MA), per backbone. Mean over 5 seeds.

Backbone	SA Acc	MA Acc	SA Hall	MA Hall
LLaVA-1.5-13B (primary)	74.3	84.8	18.7	8.9
LLaVA-Next-34B	76.1	86.3	17.2	8.4
Qwen2-VL-7B-Instruct	71.8	81.5	21.6	11.3
GPT-4V	80.5	89.2	14.0	6.4
Gemini-Pro-Vision	78.9	87.6	15.5	7.2

5.6 Ablation Study

Table 8 examines the contribution of the factual verification stage and the structured intermediate schema. When Agent 3 is removed, risk accuracy decreases from 84.8% to 79.2%, and macro-F1 drops from 83.1% to 77.8%. Grounding precision declines from 90.6% to 84.1%, and the hallucination rate increases from 8.9% to 14.8%. When the structured schema is removed (but staged prompting is kept), risk accuracy decreases to 76.8% and hallucination rate rises to 16.1%. Stage decomposition alone is therefore insufficient: the format used to pass information between stages also matters.

Table 8: Ablation study on the BDD100K benchmark (mean ± std over 5 seeds).

Variant	Risk Acc. (%)	Macro-F1 (%)	Grounding Prec. (%)	Hallucination Rate (%)	Actionability (1–5)	Latency (s)
Multi-agent w/o Agent 3	79.2 ± 0.8	77.8 ± 0.9	84.1 ± 0.6	14.8 ± 0.5	3.89 ± 0.04	4.11

(Continued)

Table 8 (continued)

Variant	Risk Acc. (%)	Macro-F1 (%)	Grounding Prec. (%)	Hallucination Rate (%)	Actionability (1–5)	Latency (s)
Multi-agent w/o structured schema	76.8 ± 0.9	75.1 ± 1.0	83.6 ± 0.6	16.1 ± 0.5	3.77 ± 0.04	3.95
Multi-agent w/o Agent 1	84.5 ± 0.7	82.8 ± 0.8	90.4 ± 0.5	9.0 ± 0.4	4.27 ± 0.03	4.62
Proposed full model	84.8 ± 0.6	83.1 ± 0.7	90.6 ± 0.5	8.9 ± 0.4	4.28 ± 0.03	4.97

5.7 Per-Agent Latency Breakdown

Table 9 reports the per-agent latency on the primary backbone (LLaVA-1.5-13B, 1×A100-80GB). Agent 2 (visual interpretation) is the dominant cost at 2.43 s per frame because it processes the full image and emits the longest free-form description; Agent 4 (risk + recommendation) is the second-largest at 1.62 s. Agent 1 (text-only relevance screening) is the cheapest at 0.18 s and could be skipped when the upstream system already restricts inputs to driving queries (the ablation in Table 8 shows that this saves 0.35 s without a meaningful accuracy loss).

Table 9: Per-agent latency on the primary configuration (s/frame, mean over the 600-frame benchmark, single A100-80GB).

Stage	Latency (s)
Agent 1—Context relevance (text-only)	0.18
Agent 2—Visual interpretation	2.43
Agent 3—Factual verification + schema	0.74
Agent 4—Risk + recommendation	1.62
Total	4.97

Deployment Positioning vs. ADAS Latency Requirements

ISO 26262 [68] and SAE J3016 [69] do not prescribe a universal latency budget for advisory ADAS. We therefore report latency as a deployment characterization rather than as evidence of real-time readiness. The 4.97 s/frame budget of the full pipeline is *not* acceptable for closed-loop control; it should be interpreted as prototype latency for off-path advisory, explanation, operator-facing, or post-event review use. Practical deployment paths to reduce latency include: (i) hierarchical invocation—run Agent 1 always, run Agents 2–4 only when the scene-complexity score (a cheap upstream signal) exceeds a threshold; (ii) distillation of Agent 2 into a smaller captioner [40,42,44]; and (iii) asynchronous execution—begin Agent 2 on frame t while Agent 4 finalizes the recommendation for frame $t-1$. We discuss these paths further in Section 7.

5.8 Confidence Calibration and Selective Abstention

We instructed Agent 4 to additionally emit a *verbalized confidence* in {low, medium, high} alongside the risk label, following the verbalized-confidence elicitation protocol of [70]. Table 10 reports ECE [64,65]

and the effect of selective abstention: when the system abstains on the bottom-15% confidence quantile and defers those frames to a human reviewer, risk accuracy on the retained 85% rises to 90.4%. The proposed multi-agent framework is also better calibrated than the single-agent baseline (ECE 0.071 vs. 0.142), which is consistent with the hypothesis that removing unsupported entities at Agent 3 produces a tighter mapping between verbalized confidence and actual correctness.

Table 10: Confidence-calibration analysis. ECE is computed with three equal-width bins on the verbalized-confidence outputs; "Acc@retained" is the risk accuracy on the 85% of frames not abstained on.

Method	ECE	Acc@retained 85% (%)
Single-agent MLLM	0.142	79.6
Proposed multi-agent	0.071	90.4

5.9 External Comparison with Published Driving-Language Systems

For external comparability we evaluated three published systems—DriveLM [22], DriveGPT4 [23], and DiLu [55]—on the risk-classification subset of our benchmark using their released checkpoints and the closest applicable prompt. Two methodological caveats apply: (i) these systems were not trained on a 3-class {low, medium, high} risk-level objective; we therefore mapped their native outputs to the 3-class label using a deterministic rule documented in Appendix B, and (ii) DriveLM expects a graph-VQA input, which we approximated by serializing our schema into its expected format. Table 11 reports the comparison. The proposed framework outperforms all three on risk accuracy and hallucination rate; we explicitly do *not* claim a like-for-like benchmark and treat these numbers as external sanity checks rather than head-to-head winners.

Table 11: External comparison with published driving-language systems on the risk-classification subset of our benchmark. See text for methodological caveats.

System	Risk Acc. (%)	Hallucination (%)
DriveLM [22]	71.5	19.4
DriveGPT4 [23]	76.2	16.1
DiLu [55]	72.8	18.7
Proposed multi-agent	84.8	8.9

5.10 Human Evaluation

Three blinded raters from the evaluation panel (Section 4.6) further evaluated the final recommendations in terms of usefulness, specificity, and consistency with visible evidence. The proposed framework achieves an average actionability score of 4.28/5 (ICC(2, k) = 0.74), compared with 3.62/5 for the single-agent baseline. The raters also noted that the outputs produced by the multi-agent framework were generally less speculative and easier to audit, which is consistent with the lower hallucination rate in Table 2.

6 Qualitative Analysis

6.1 Stepwise Per-Agent Intermediate Outputs

Table 12 traces the four-agent pipeline on a representative pedestrian–vehicle interaction frame from the benchmark (frame ID `bdd100k_val_b1d10d08-743cd5d9_017`). The single-agent baseline

mentions a "cyclist about to cross," which is not visible in the frame; the proposed Agent 3 schema correctly drops this entity (its `visually_supported` flag is `false`), so it does not enter the recommendation produced by Agent 4. Two additional worked examples (a sign-interpretation scene and a nighttime density scene) are provided in Appendix A.

Table 12: Stepwise intermediate outputs of the four agents on a representative pedestrian–vehicle interaction frame. Compare with the single-agent baseline output at the bottom.

Stage	Output (Excerpt)
Agent 1 (relevance)	`{relevance: relevant, query_type: pedestrian_interaction}`
Agent 2 (visual)	"Two-lane urban street, daytime, light rain. A pedestrian in dark clothing stands at the right curb; a parked van is on the right shoulder. A traffic sign on the left appears to be a no-parking sign. A cyclist may be approaching from the right but is hard to confirm."
Agent 3 (verified schema)	`{sign_intersection: [{type: no_parking, applicability: right_shoulder, visually_supported: true}], traffic_density: [{lane: left, density: sparse}, {lane: right, density: sparse, ego_distance: short}], vulnerable_road_users: [{class: pedestrian, location: right_curb, motion_state: standing, visually_supported: true}], temporary_road_conditions: [{type: parked_van, location: right_shoulder, blocking_status: partial, visually_supported: true}], attention_points: [{cue: cyclist_from_right, visually_supported: false}]}`
Agent 4 (risk + advice)	"Risk: **medium**. Reasons: pedestrian standing at the right curb in light rain; right-shoulder parked van partially blocks lateral visibility. Recommendation: reduce speed, increase lateral clearance from the right curb, and be prepared for the pedestrian to enter the road."
Single-agent baseline (same frame)	"Risk: medium. A pedestrian is at the curb **and a cyclist is about to cross**; recommend slow down and yield to the cyclist." (*bolded phrase: hallucinated entity removed by Agent 3 in our pipeline.*)

6.2 Failure-Case Analysis

Table 13 reports two representative failure modes of the proposed framework. (F1) On a heavily occluded nighttime pedestrian scene, Agent 2 missed the pedestrian entirely (low descriptive recall), so Agent 3 had nothing to verify and Agent 4 issued a "low-risk" label that the ground truth marked as "high." (F2) On a complex five-way intersection with two simultaneously valid traffic-control cues (a temporary stop sign held by a flagger and an active green signal), Agent 3 retained both in its schema but Agent 4 prioritized the green

signal over the flagger and issued a "proceed with caution" advisory instead of the correct "stop and yield to flagger." These failure modes share a common cause: Agent 3 cannot recover information that was never produced by Agent 2, and Agent 4 has no principled mechanism for resolving conflicts between two visually supported cues. Both motivate the future-work directions in Section 7.1.

Table 13: Representative failure cases of the proposed framework.

ID	Failure Description
F1	Heavy nighttime occlusion; Agent 2 missed the pedestrian entirely. Predicted "low," ground truth "high." Root cause: descriptive recall at Agent 2; Agent 3 cannot recover what Agent 2 did not emit.
F2	Conflicting traffic-control cues (temporary stop sign held by flagger vs. active green signal). Agent 3 retained both, Agent 4 picked the wrong one. Root cause: lack of priority arbitration in the recommendation prompt.

7 Discussion

The empirical pattern is consistent across the main table, the scenario split, the per-class analysis, the ablation study, the backbone stress test, and the robustness analysis: the framework gains most when it prevents weakly grounded descriptions from directly driving the final recommendation. That observation helps explain why the method improves both prediction quality and grounding quality. A single-agent response can be globally plausible while still mixing observable facts with inferred details that have not been visually established. Once those two layers are fused, later auditing becomes difficult. The multi-agent design interrupts that process.

Agent 2 and Agent 3 serve different functions. Agent 2 favors descriptive recall and tries not to miss relevant cues. Agent 3 then performs the opposite operation. It compresses the raw description, discards unsupported or decision-irrelevant statements, and normalizes the retained evidence into a typed JSON schema. That separation matters. In a single monolithic prompt, missing a cue and over-interpreting a cue are hard to disentangle because both errors appear in the same free-form output. In the proposed design, they occur at different stages and can therefore be inspected separately. Failure case F1 in Table 13 shows what happens when Agent 2 fails at recall: Agent 3 cannot recover what was never produced. Failure case F2 shows what happens when Agent 3 retains conflicting cues without priority guidance: Agent 4 may pick the wrong one. Both failure modes point to specific, addressable design improvements rather than to a fundamental limit of the approach.

This is also why the ablation results are informative. Removing Agent 3 degrades grounding precision and raises hallucination rate more sharply than removing the structured schema alone. The framework still benefits from staged prompting without the schema, but it loses a meaningful portion of the gain once the verified intermediate representation disappears. The mechanism is therefore not prompt count; it is control over what information is allowed to survive into the final risk reasoning stage. The backbone stress test (Section 5.5) reinforces this conclusion: the relative gain is consistent across LLaVA-1.5-13B, LLaVA-Next-34B, Qwen2-VL-7B-Instruct, GPT-4V, and Gemini Pro Vision. The framework therefore behaves as an architectural property of the inference pipeline rather than as a side-effect of a particular backbone.

Pedestrian-related scenes show the largest improvement because they concentrate precisely the type of evidence that monolithic prompting handles poorly: local, safety-critical, and easily overstated cues. The per-class analysis (Table 4) and confusion matrix (Table 5) show that the gain is broadly distributed across

the three risk classes, with the largest absolute lift on the under-represented "high" class. This is the most practically useful direction in which to push performance, because false-low predictions on actually-high-risk frames are the dominant safety-relevant failure mode.

The calibration analysis (Section 5.8) adds a complementary point. Better grounding tightens the mapping between verbalized confidence and actual correctness (ECE drops from 0.142 to 0.071), which makes selective abstention more useful: when the system defers the bottom-15% confidence quantile to a human reviewer, accuracy on the retained frames rises to 90.4%. This indicates that the framework is not only more accurate on average but also more honest about when it is unsure—a property that matters for real ADAS deployment where unreliable predictions must be flagged rather than acted upon.

The price is latency. A multi-stage reasoning chain is slower than a single prompt. The per-agent breakdown (Table 9) localizes the cost: Agent 2 (visual interpretation) is the dominant stage, followed by Agent 4. For closed-loop vehicle control, the 4.97 s/frame budget is not defensible. For explainable ADAS modules, operator-facing interfaces, post-event review, or offline safety analysis, the trade-off is more reasonable because transparency and auditability are part of the product requirement rather than secondary features. Three concrete optimization paths follow directly from the breakdown: (i) hierarchical invocation (run Agent 1 always; run Agents 2–4 only on complex scenes); (ii) distillation of Agent 2 into a smaller captioner [40,42,44]; (iii) asynchronous execution across consecutive frames. The framework is therefore best understood as a semantic support layer that complements perception and planning rather than replacing them.

External comparison with DriveLM, DriveGPT4, and DiLu (Table 11) places the framework above these published systems on our risk benchmark, but with explicit caveats: the published systems were not trained on a 3-class risk-level objective and we had to map their native outputs. We deliberately do *not* claim a like-for-like benchmark win and instead treat this comparison as an external sanity check that the proposed gains are not an artifact of evaluating only against in-house baselines.

7.1 Limitations and Future Directions

Several limitations remain. First, the current evaluation is open-loop and frame-centric. The benchmark contains 600 manually labeled BDD100K frames; that is sufficient for a controlled comparison of reasoning structure, but it cannot determine whether the generated recommendations improve downstream driving behavior in temporally evolving traffic situations. Frame-level evaluation also cannot capture motion continuity, intent development, or evidence accumulation across adjacent frames.

Second, the framework still inherits the weaknesses of the underlying multimodal model. Performance declines under nighttime, rain/fog, and occlusion even though the proposed design degrades more gracefully than the single-agent baseline. The calibration analysis (Section 5.8) is a first step toward principled handling of this residual uncertainty, but more work is needed—including semantic-uncertainty estimation [71], Bayesian-style aleatoric/epistemic decomposition [72], and confidence-aware selective abstention thresholds tuned per scene condition.

Third, the empirical comparison, although broader than in the original submission, is not exhaustive. We have added chain-of-thought, ReAct-style, and three published driving-language baselines (DriveLM, DriveGPT4, DiLu), but a fully like-for-like comparison would require re-training several of these systems on our risk objective, which is left to future work.

These limitations also define the next research agenda. A natural extension is to move from frame-level reasoning to short video clips and temporally consistent scene interpretation, which would let Agent 3 accumulate evidence across frames and Agent 4 reason about intent rather than instantaneous state. A

second is to incorporate principled confidence calibration and selective abstention into the recommendation policy, allowing the system to indicate when the visible evidence is insufficient. A third is deployment-oriented optimization: hierarchical invocation, distillation of Agent 2 into a lightweight captioner [40,42,44], and asynchronous cross-frame execution, with explicit latency targets aligned with the intended ADAS profile [68,69]. A fourth is closed-loop validation in CARLA or in a vehicle-in-the-loop setup, which would close the gap between our open-loop semantic evaluation and end-to-end driving safety.

8 Conclusion

This paper presented a grounded multi-agent multimodal LLM framework for interpretable risk assessment in driving scenes. We decomposed the reasoning process into four sequential agents—context relevance, visual interpretation, factual verification with anomaly extraction, and risk assessment with recommendation—and inserted a typed intermediate schema that prevented unsupported content from reaching the final advisory. We constructed a 600-frame BDD100K benchmark with three-rater annotation, evaluated the framework with mean ± std over five seeds, and confirmed the gains with McNemar tests; on this benchmark, the framework improved risk accuracy from 74.3% to 84.8%, raised macro-F1 from 72.8% to 83.1%, cut the hallucination rate from 18.7% to 8.9%, and lifted the actionability score from 3.62 to 4.28. We additionally validated the framework on five MLLM backbones (LLaVA-1.5-13B, LLaVA-Next-34B, Qwen2-VL-7B-Instruct, GPT-4V, and Gemini Pro Vision), profiled per-agent latency, analyzed confidence calibration with selective abstention, compared against three published driving-language systems, and provided a failure-case analysis on the residual hard frames. The full agent prompts, the annotation guideline, the actionability and hallucination rubrics, and the per-condition frame counts are released in the Appendix to support independent replication.

Looking forward, three concrete directions should be prioritized in subsequent work. (1) *Temporal extension and closed-loop validation*. Extend the framework from single frames to short video clips so that Agent 3 can accumulate evidence across frames, and validate the resulting advisories in CARLA or in a vehicle-in-the-loop setup. (2) *Latency-aware deployment*. Pursue hierarchical invocation (run Agents 2–4 only on complex scenes), distill Agent 2 into a lightweight captioner, and adopt asynchronous cross-frame execution, with explicit latency targets aligned with the intended ADAS profile [68,69]. (3) *Principled uncertainty handling*. Replace the current verbalized-confidence proxy with semantic uncertainty estimation [71] and Bayesian aleatoric/epistemic decomposition [72], and tune the selective-abstention threshold per scene condition. These directions, combined with the appendix-based release of the prompts, annotation guidelines, rubrics, and per-condition frame counts, can move language-grounded driving systems from controlled benchmarks toward deployable, auditable decision-support layers for real ADAS products.

Acknowledgement: The authors thank the National Center for High-Performance Computing (NCHC) for providing computational resources and technical support.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Conceptualization, Chien-Hao Tseng and Chung-I Huang; methodology, Chien-Hao Tseng, Min-Yu Chen, Meng-Wei Lin and Chung-I Huang; validation, Chien-Hao Tseng, Min-Yu Chen, Meng-Wei Lin and Jyh-Horng Wu; formal analysis, Chien-Hao Tseng and Chung-I Huang; investigation, Chien-Hao Tseng, Min-Yu Chen and Meng-Wei Lin; writing—original draft preparation, Chien-Hao Tseng and Chung-I Huang; writing—review and editing, all authors; supervision, Jyh-Horng Wu and Chung-I Huang. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The public source data used in this study are available from the BDD100K dataset. The data and instrumentation used for the statistical analysis are not publicly deposited in an online repository because they include project-specific annotations and evaluation records prepared for this study. De-identified statistical summaries, analysis protocols, and implementation details are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A Full Agent Prompts and Worked Example

Appendix A.1 Agent 1—Context Relevance Evaluation (Text-Only)

```
SYSTEM: You are a relevance classifier for a driving-scene assistant.
USER:   Given the user request below, output ONE token from
        {relevant, partial, irrelevant} and a short query_type.
        Output JSON only: {''relevance'': ..., ''query_type'': ...}.
        REQUEST: <user request>
```

Appendix A.2 Agent 2—Visual Perception and Interpretation

```
SYSTEM: You are a careful driving-scene observer. Describe what is
        VISIBLE in the image. Cover: road layout, lanes, traffic
        participants, traffic signs/signals, lighting, weather, road
        condition, salient hazards. Prefer broad recall; downstream
        stages will filter unsupported details. Do NOT infer intent.
USER:   <image> Describe the scene.
```

Appendix A.3 Agent 3—Factual Verification and Anomaly Extraction

```
SYSTEM: You are a factual verifier. Given (a) the image and (b) the
        Agent-2 description, extract ONLY content that is visually
        supported in the image. For each item set
        ''visually_supported'': true|false. Output JSON ONLY in this
        schema:
        {
          ''sign_intersection'': [{type, content, applicability,
                                   visually_supported}],
          ''traffic_density'':   [{lane, density, ego_distance,
                                   visually_supported}],
          ''vulnerable_road_users'':
                                 [{class, location, motion_state,
                                   visually_supported}],
          ''temporary_road_conditions'':
                                 [{type, location, blocking_status,
                                   visually_supported}],
          ''attention_points'':  [{cue, visually_supported}]
        }
        Items with visually_supported=false will be DROPPED.
```

```
USER:   <image> + Agent-2 output above.
```

Appendix A.4 Agent 4—Risk Assessment and Recommendation

```
SYSTEM: You are a driving-risk advisor. Use ONLY the verified schema
        below (do NOT use the raw caption). Output:
        {
          ''risk_level'':  one of {low, medium, high},
          ''confidence'':  one of {low, medium, high},
          ''reasons'':     list of short evidence-grounded reasons,
          ''recommendation'': stepwise advisory text
        }
        Policy: low-risk -> concise confirmation;
                medium/high-risk -> stepwise advisory.
USER:   <verified schema from Agent 3>
```

Appendix A.5 API-Specific Deviations

For GPT-4V (`gpt-4-turbo-2024-04-09`) and Gemini Pro Vision, image inputs are passed as base64 PNG instead of via the LLaVA preprocessor; sampling defaults to the closest supported parameters (`temperature = 0.2`, `top_p = 0.9`, `max_output_tokens` matching the per-agent values listed in Section 3.7).

Appendix B Annotation Guideline (Excerpt)

Risk-Level Decision Rule

High: at least one of: vulnerable road user in or about to enter the ego trajectory; explicit stop/yield obligation that the ego is approaching; a hazard that, if ignored within 2 s, would cause a near-miss. **Medium**: a salient cue requires deliberate driver action (brake, slow, change lane) but no immediate near-miss is implied. **Low**: routine driving conditions; no cue requires deliberate corrective action beyond normal lane keeping.

Appendix B.1 Disagreement Resolution

Three independent raters; if two agree, the majority label wins. If 1-1-1 (3.5% of frames), a senior fourth rater adjudicates after reviewing the three rationales.

Appendix B.1.1 External-Baseline Output Mapping (Used in Section 5.9)

DriveLM/DriveGPT4 free-text outputs are mapped to {low, medium, high} using a fixed keyword list (e.g., "proceed normally" → low; "slow down," "yield," "be cautious" → medium; "stop," "brake hard," "avoid" → high). DiLu's discrete decisions are mapped using its native action set (no-op → low; lane-change/decelerate → medium; emergency-brake → high).

Appendix C Hallucination Scoring Rubric

A frame is counted as a hallucinated output if its recommendation contains *any* of the three categories below. The decision is binary; the granular per-category counts are kept for diagnostic purposes only.

- **H1—Invented entity.** The output asserts a vehicle, person, sign, lane, or signal that is not visible in the frame.

- **H2—Contradicted attribute.** An asserted attribute conflicts with a visible attribute (e.g., "red light" vs. visibly green signal).
- **H3—Invented relation.** A spatial or interaction relation between two entities is asserted without visual support (e.g., "the cyclist is crossing in front of the ego" when the cyclist is parallel to the ego on the sidewalk).

Partial visual support (e.g., "a pedestrian may be present") is *not* hallucination as long as no contradictory or invented entity/attribute/relation is asserted.

Appendix D Per-Condition Frame Counts and Supplementary Tables

Per-condition frame counts for Table 6 are reported in the main text alongside the table. We additionally release: (i) the per-frame predicted-label CSV for all five seeds and all methods, (ii) the verbalized-confidence outputs used in the calibration analysis, and (iii) the blinded human-evaluation scoring sheet. These supplementary artifacts are kept compact: The large diagnostic dumps that appeared in the original Round-1 appendix have been removed in this revision, in line with the reviewers' request.

References

1. Yurtsever E, Lambert J, Carballo A, Takeda K. A survey of autonomous driving: common practices and emerging technologies. IEEE Access. 2020;8:58443–69.
2. Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: the KITTI dataset. Int J Robot Res. 2013;32(11):1231–7.
3. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, et al. The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 3213–23.
4. Yu F, Chen H, Wang X, Xian W, Chen Y, Liu F, et al. BDD100K: a diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 2636–45.
5. Caesar H, Bankiti V, Lang AH, Vora S, Liong VE, Xu Q, et al. nuScenes: a multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 11621–31.
6. Sun P, Kretzschmar H, Dotiwalla X, Chouard A, Patnaik V, Tsui P, et al. Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 2446–54.
7. Liao H, Li Y, Li ZN, Bian Z, Lee J, Cui Z, et al. Real-time accident anticipation for autonomous driving through monocular depth-enhanced 3D modeling. Accid Anal Prev. 2024;207(7):107760. doi:10.1016/j.aap.2024.107760.
8. Grosse K, Alahi A. A qualitative AI security risk assessment of autonomous vehicles. Transp Res Part C Emerg Technol. 2024;169(11):104797. doi:10.1016/j.trc.2024.104797.
9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:6000–10. doi:10.65215/pc26a033.
10. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901.
11. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 2022;35:27730–44. doi:10.52202/068431-2011.
12. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24; Virtual. p. 8748–63.
13. Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. Flamingo: a visual language model for few-shot learning. Adv Neural Inf Process Syst. 2022;35:23716–36.

14. Li J, Li D, Savarese S, Hoi SCH. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: Proceedings of the 40th International Conference on Machine Learning; 2023 Jul 23–29; Honolulu, HI, USA. p. 19730–42.
15. Zhou X, Liu M, Yurtsever E, Zagar BL, Zimmer W, Cao H, et al. Vision language models in autonomous driving: a survey and outlook. IEEE Trans Intell Veh. 2024. doi:10.1109/tiv.2024.3402136.
16. Gan L, Chu W, Li G, Tang X, Li K. Large models for intelligent transportation systems and autonomous vehicles: a survey. Adv Eng Inform. 2024;62:102786. doi:10.1016/j.aei.2024.102786.
17. Cui C, Ma Y, Cao X, Ye W, Zhou Y, Liang K, et al. A survey on multimodal large language models for autonomous driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops; 2024 Jan 3–8; Waikoloa, HI, USA. p. 958–79.
18. Park S, Lee M, Kang J, Choi H, Park Y, Cho J, et al. VLAAD: vision and language assistant for autonomous driving. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops; 2024 Jan 3–8; Waikoloa, HI, USA. p. 980–7.
19. Inoue Y, Yada Y, Tanahashi K, Yamaguchi Y. NuScenes-MQA: integrated evaluation of captions and QA for autonomous driving datasets using markup annotations. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops; 2024 Jan 3–8; Waikoloa, HI, USA. p. 930–8.
20. Cao X, Zhou T, Ma Y, Ye W, Cui C, Tang K, et al. MAPLM: a real-world large-scale vision-language benchmark for map and traffic scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jan 3–8; Waikoloa, HI, USA. p. 21819–30.
21. Shao H, Hu Y, Wang L, Waslander SL, Liu Y, Li H. LMDrive: closed-loop end-to-end driving with large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jan 3–8; Waikoloa, HI, USA. p. 15120–30.
22. Sima C, Renz K, Chitta K, Chen L, Zhang H, Xie C, et al. DriveLM: driving with graph visual question answering. In: Proceedings of the Computer Vision—ECCV 2024; 2024 Sep 29–Oct 4; Milan, Italy. p. 256–74.
23. Xu Z, Zhang Y, Xie E, Zhao Z, Guo Y, Wong KYK, et al. DriveGPT4: interpretable end-to-end autonomous driving via large language model. IEEE Robot Autom Lett. 2024;9(10):8186–93.
24. Rohrbach A, Hendricks LA, Burns K, Darrell T, Saenko K. Object hallucination in image captioning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018 Oct 31–Nov 4; Brussels, Belgium. p. 4035–45.
25. Li Y, Du Y, Zhou K, Wang J, Zhao WX, Wen JR. Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023 Dec 6–10; Singapore. p. 292–305.
26. Li J, Cheng X, Zhao WX, Nie JY, Wen JR. HaluEval: a large-scale hallucination evaluation benchmark for large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing; 2023 Dec 6–10; Singapore. p. 6449–64.
27. Yao S, Zhao J, Yu D, Du N, Shafran I, Narasimhan K, et al. ReAct: synergizing reasoning and acting in language models. In: Proceedings of the 11th International Conference on Learning Representations (ICLR); 2023 May 1–5; Kigali, Rwanda.
28. Dosovitskiy A, Ros G, Codevilla F, López AM, Koltun V. CARLA: an open urban driving simulator. In: Proceedings of the 1st Annual Conference on Robot Learning; 2017 Nov 13–15; Mountain View, CA, USA. p. 1–16.
29. Demmel S, Glaser S, Gruyer D, Larue G, Orfila O, Rakotonirainy A, et al. Global risk assessment in an autonomous driving context: impact on both the car and the driver. IFAC-PapersOnLine. 2019;51(34):390–5.
30. Erdoğan M, Kaya İ., Karaşan A, Çolak M. Evaluation of autonomous vehicle driving systems for risk assessment based on three-dimensional uncertain linguistic variables. Appl Soft Comput. 2021;113(2):107934. doi:10.1016/j.asoc.2021.107934.
31. Yang Z, Jia X, Li H, Yan J. LLM4Drive: a survey of large language models for autonomous driving. arXiv:2311.01043. 2023.

32. Chang X, Xue M, Liu X, Pan Z, Wei X. Driving by the rules: a benchmark for integrating traffic sign regulations into vectorized HD map. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2025 Jun 11–15; Nashville, TN, USA. p. 6823–33.
33. Tian K, Mao J, Zhang Y, Jiang J, Zhou Y, Tu Z. NuScenes-SpatialQA: a spatial understanding and reasoning benchmark for vision-language models in autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops; 2025 Oct 19–20; Honolulu, HI, USA. p. 4626–35.
34. Xu Z, Bai Y, Zhang Y, Li Z, Xia F, Wong KYK, et al. DriveGPT4-V2: harnessing large language model capabilities for enhanced closed-loop autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2025 Jun 11–15; Nashville, TN, USA. p. 17261–70.
35. Renz K, Chen L, Arani E, Sinavski O. SimLingo: vision-only closed-loop autonomous driving with language-action alignment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2025 Jun 11–15; Nashville, TN, USA. p. 11993–2003.
36. Fu H, Zhang D, Zhao Z, Cui J, Liang D, Zhang C, et al. ORION: a holistic end-to-end autonomous driving framework by vision-language instructed action generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2025 Oct 19–23; Honolulu, HI, USA. p. 24823–34.
37. Wang S, Yu Z, Jiang X, Lan S, Shi M, Chang N, et al. OmniDrive: a holistic vision-language dataset for autonomous driving with counterfactual reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2025 Jun 11–15; Nashville, TN, USA. p. 22442–52.
38. Lu Y, Tu J, Ma Y, Zhu X. ReAL-AD: towards human-like reasoning in end-to-end autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2025 Oct 19–23; Honolulu, HI, USA. p. 27783–93.
39. Tian X, Gu J, Li B, Liu Y, Wang Y, Zhao Z, et al. DriveVLM: the convergence of autonomous driving and large vision-language models. arXiv:2402.12289. 2024.
40. Qiao Z, Li H, Cao Z, Liu HX. LightEMMA: lightweight end-to-end multimodal model for autonomous driving. arXiv:2505.00284. 2025.
41. Tian K, Zhou Y, Tu Z. OpenEMMA: open-source multimodal model for end-to-end autonomous driving. arXiv:241215208. 2024.
42. Zhang R, Zhang W, Tan X, Yang S, Wan X, Luo X, et al. VLDrive: vision-augmented lightweight MLLMs for efficient language-grounded autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2025 Oct 19–23; Honolulu, HI, USA. p. 5923–33.
43. Xie Y, Xu R, He T, Hwang JJ, Luo K, Ji J, et al. S4-Driver: scalable self-supervised driving multimodal large language model with spatio-temporal visual representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2025 Jun 11–15; Nashville, TN, USA. p. 1622–32.
44. Hegde D, Yasarla R, Cai H, Han S, Bhattacharyya A, Mahajan S, et al. Distilling multi-modal large language models for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2025 Jun 11–15; Nashville, TN, USA. p. 27575–85.
45. Zhang Z, Li X, Xu Z, Peng W, Zhou Z, Shi M, et al. MPDrive: improving spatial understanding with marker-based prompt learning for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2025 Jun 11–15; Nashville, TN, USA. p. 12089–99.
46. L'ubberstedt J, Guerrero ER, Uhlemann N, Lienkamp M. V3LMA: visual 3D-enhanced language model for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2025 Jun 11–15; Nashville, TN, USA. p. 4808–17.
47. Carrasco S, Fernández Llorca D, Sotelo MÁ. SCOUT: Socially-COnsistent and UndersTandable graph attention network for trajectory prediction of vehicles and VRUs. arXiv:2102.06361. 2021.
48. Lorenzo J, Parra I, Sotelo MÁ. IntFormer: predicting pedestrian intention with the aid of the transformer architecture. arXiv:2105.08647. 2021.
49. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst. 2022;35:24824–37. doi:10.59350/bkm9q-11k47.

50. Schick T, Dwivedi-Yu J, Dessì R, Raileanu R, Lomeli M, Hambro E, et al. Toolformer: language models can teach themselves to use tools. Adv Neural Inf Process Syst. 2023;36:68539–51.
51. Shinn N, Cassano F, Gopinath A, Narasimhan K, Yao S. Reflexion: language agents with verbal reinforcement learning. Adv Neural Inf Process Syst. 2023;36:8634–52. doi:10.52202/075280-0377.
52. Wang L, Ma C, Feng X, Zhang Z, Yang H, Zhang J, et al. A survey on large language model based autonomous agents. Front Comput Sci. 2024;18(6):186345. doi:10.1007/s11704-024-40231-1.
53. Durante Z, Huang Q, Wake N, Gong R, Park JS, Sarkar B, et al. Agent AI: surveying the horizons of multimodal interaction. arXiv:2401.03568. 2024.
54. Fu D, Li X, Wen L, Dou M, Cai P, Shi B, et al. Drive like a human: rethinking autonomous driving with large language models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops; 2024 Jan 3–8; Waikoloa, HI, USA. p. 910–9.
55. Wen L, Fu D, Li X, Cai X, Ma T, Cai P, et al. DiLu: a knowledge-driven approach to autonomous driving with large language models. In: Proceedings of the 12th International Conference on Learning Representations (ICLR); 2024 May 7–11; Vienna, Austria.
56. Sha H, Mu Y, Jiang Y, Chen L, Xu C, Luo P, et al. LanguageMPC: large language models as decision makers for autonomous driving. arXiv:2310.03026. 2023.
57. Mao J, Ye J, Qian Y, Pavone M, Wang Y. A language agent for autonomous driving. arXiv:2311.10813. 2023.
58. Powers DMW. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. J Mach Learn Technol. 2011;2(1):37–63.
59. Liu H, Li C, Li Y, Lee YJ. Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2024 Jun 16–22; Seattle, WA, USA. p. 26296–306.
60. Wang P, Bai S, Tan S, Wang S, Fan Z, Bai J, et al. Qwen2-VL: enhancing vision-language model's perception of the world at any resolution. arXiv:2409.12191. 2024.
61. OpenAI. GPT-4V(ision) system card. OpenAI System Card. 2023.
62. Gemini Team, Google. Gemini: a family of highly capable multimodal models. arXiv:2312.11805. 2023.
63. Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971;76(5):378–82. doi:10.1037/h0031619.
64. Naeini MP, Cooper GF, Hauskrecht M. Obtaining well calibrated probabilities using bayesian binning. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2015 Jun 25–30; Austin, TX, USA. p. 2901–7.
65. Guo C, Pleiss G, Sun Y, Weinberger KQ. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning; 2017 Aug 6–11; Sydney, Australia. p. 1321–30.
66. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153–7. doi:10.1007/bf02295996.
67. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Comput. 1998;10(7):1895–923. doi:10.1162/089976698300017197.
68. ISO Standard 26262. Road vehicles—functional safety. Geneva, Switzerland: International Organization for Standardization; 2018.
69. SAE International Standard J3016. SAE J3016: taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Warrendale, PA, USA: SAE International; 2021.
70. Xiong M, Hu Z, Lu X, Li Y, Fu J, He J, et al. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In: Proceedings of the 12th International Conference on Learning Representations (ICLR); 2024 May 7–11; Vienna, Austria.
71. Kuhn L, Gal Y, Farquhar S. Semantic uncertainty: linguistic invariances for uncertainty estimation in natural language generation. In: Proceedings of the 11th International Conference on Learning Representations (ICLR); 2023 May 1–5; Kigali, Rwanda.
72. Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? Adv Neural Inf Process Syst. 2017;30:5574–84.