CoherentDrive: Conflict-Aware World-State Grounding for Hallucination-Resistant Autonomous Driving Reasoning

Shuo Liu[a], Lei Shi[b], Jing Xu[a], Yufei Gao[b], Yucheng Shi[a*]
[a] School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, China
[b] School of Cyber Science and Engineering, Zhengzhou University, Zhengzhou, China
[*] Corresponding author
lius@gs.zzu.edu.cn, ieycshi@zzu.edu.cn

***Abstract*—Large language models (LLMs) are increasingly used to support language-based interaction in autonomous driving, including scene question answering and risk explanation. Yet a critical reliability failure persists: when prompted with long, redundant, and mutually inconsistent perception logs from heterogeneous sensors, LLMs may produce fluent but unsupported statements, hallucinate non-existent traffic agents, and amplify false premises. We address this problem from an interface perspective. We present *CoherentDrive*, a prompt-ready grounding framework that compiles multi-source perception facts into a compact, conflict-resolved *Coherent World State (CWS)* with entity identifiers, provenance, and uncertainty flags. CoherentDrive further enforces a grounding-first answering protocol that restricts responses to CWS-supported entities and relations, enabling explicit refusal when evidence is missing. On nuScenes-QA under a zero-shot setting with multiple LLM backbones, CoherentDrive consistently improves answer correctness while substantially reducing hallucinated entity mentions, and in a single-pass CWS prompting setting it also shortens prompt length and reduces inference latency. Extensive experiments demonstrate that conflict-resolved world-state prompting provides an effective and reliable foundation for hallucination-resistant driving question answering.**

***Index Terms*—Autonomous driving, Large language models, Hallucination mitigation, World-state compilation.**

I. Introduction

Recent progress in autonomous driving perception is driven by large-scale benchmarks such as nuScenes [1] and the Waymo Open Dataset [2] and by bird's-eye-view (BEV) representations that align heterogeneous sensors into a unified spatial frame. BEV-centric perception and fusion methods have achieved strong 3D scene understanding performance across camera-only and multi-sensor settings [3], [4], [5], [6]. In parallel, large multimodal models and vision-language models (VLMs) have enabled language-conditioned perception and reasoning [7], [8], [9]. These advances motivate language-grounded driving benchmarks that probe scene-level understanding beyond detection, such as nuScenes-QA [10] and graph-based driving VQA tasks like DriveLM [11]. Recent work also highlights the importance of aligning language understanding with closed-loop driving actions [12].

However, reliability remains a major barrier when deploying LLM/VLM reasoning for driving-critical interactions. Hallucination—fluent yet unsupported generation—is widely recognized as a core failure mode in natural language generation and LLM systems [13], [14]. In driving question answering, hallucinations often arise not only from model limitations, but also from *evidence contamination*: prompts that concatenate raw perception outputs tend to be (i) excessively long, (ii) redundant across modalities, and (iii) internally inconsistent due to misalignment, missed detections, or conflicting attributes. When asked to answer a question under such evidence, an LLM may implicitly "reconcile" the logs by inventing entities or relations, or by accepting false premises as facts—behaviors that are unacceptable in safety-critical contexts.

A natural response is to improve prompting and reasoning strategies. Chain-of-thought prompting can enhance multi-step reasoning [15], and agentic paradigms such as ReAct and Reflexion combine reasoning with action and self-feedback [16], [17]. Meanwhile, the MLLM community has started to develop stronger hallucination evaluation and mitigation tools, such as open-set protocols (ODE [18]) and decoding-time interventions (FarSight [19]), further motivating reliability-focused driving cognition. Yet these approaches primarily change how a model reasons; they do not guarantee grounding when the input evidence is contradictory or unstructured. In other words, if the prompt does not provide a single, verifiable source of truth, stronger reasoning workflows may still propagate inconsistent evidence into confident answers.

This paper argues that hallucination-resistant driving QA should be treated as an interface design problem: before asking an LLM to reason, we should compile a compact, contradiction-aware world state that can be explicitly referenced and checked.

We introduce CoherentDrive, summarized in Fig. 1. As illustrated, directly prompting long raw perception logs can lead the LLM to accept a false premise (e.g., a non-existent bicycle) and hallucinate unsupported entities, whereas CoherentDrive compiles the logs via the *World-State Compiler (WSC)* into a compact *Coherent World State (CWS)* and applies a grounding-first protocol (cite entity IDs / refuse unsupported premises) to produce auditable answers and actions.

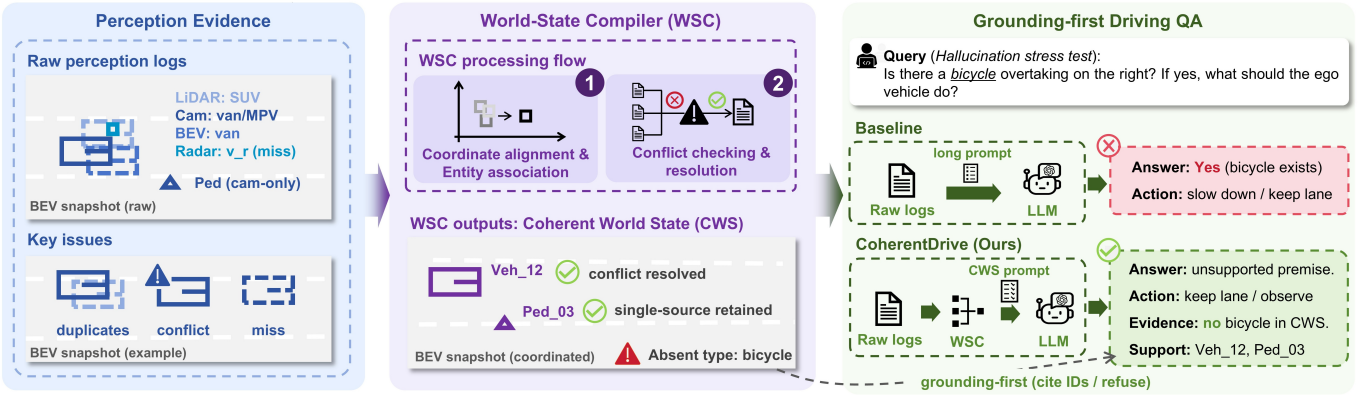


Fig. 1. CoherentDrive overview: from uncoordinated perception evidence to grounding-first driving QA. **Left:** Raw multi-sensor perception logs can be redundant, conflicting, and incomplete (duplicates/conflicts/misses), which contaminates the prompt evidence. **Middle:** The World-State Compiler (WSC) performs coordinate alignment and entity association, followed by conflict checking and resolution, to compile a Coherent World State (CWS) that preserves a single truth with provenance cues. **Right:** In a hallucination stress-test query with a false premise, using a long raw-log prompt may hallucinate a bicycle and yield unsafe recommendations, while CoherentDrive uses a compact CWS prompt and a grounding-first protocol (cite IDs / refuse) to reject unsupported premises and output auditable actions with supporting entity IDs.

Our contributions are mainly as follows:

1) We reveal that hallucinations in open-world driving QA are often triggered by evidence contamination from redundant, conflicting, and incomplete multi-source perception logs, and we cast hallucination-resistant driving QA as a coordination-first interface problem.

2) We propose CoherentDrive, a reliability-oriented prompting framework that introduces a deterministic World-State Compiler (WSC) to compile heterogeneous perception facts into a compact Coherent World State (CWS) with explicit conflict handling, provenance, and uncertainty flags, and we adopt a grounding-first answering protocol that enforces evidence-supported responses and premise rejection.

3) Extensive experiments on nuScenes-QA with multiple LLM backbones demonstrate that CoherentDrive improves QA correctness while substantially reducing hallucinated entities, and additional controlled tests further verify robustness under conflict- and premise-induced perturbations.

II. Method

A. Overview

CoherentDrive targets reliability-oriented driving question answering (Driving QA) by explicitly separating (i) *multi-source evidence coordination* from (ii) *language-based reasoning*. As illustrated in Fig. 1, heterogeneous perception agents first emit structured evidence. A deterministic World-State Compiler (WSC) module compiles these uncoordinated outputs into a conflict-resolved, provenance-aware Coherent World State (CWS). Then, a **grounding-first, verifier-aware reasoning** engine answers queries and produces audit-friendly decisions, with every factual claim explicitly grounded to entity IDs and fields in the CWS. The overall procedure is summarized in Algorithm 1.

B. Inputs, Outputs, and Notation

At each timestamp t, K perception agents output structured facts $\{\mathcal{F}_k^t\}_{k=1}^K$ in a unified interface (JSON in our implementation). Each $\mathcal{F}_k^t$ contains a set of observations $\{o_{k,i}\}$ with (optional) state mean $\mathbf{x}_{k,i}$, covariance $\mathbf{P}_{k,i}$, class distribution $p_{k,i}(c)$, confidence $q_{k,i} \in [0,1]$, and a provenance tag $\texttt{src} = k$. Given an optional user query Q, CoherentDrive outputs: (1) a structured decision D_{final} (JSON), e.g., `recommended_action`, `confidence`, `supporting_entity_ids`; and (2) a natural-language justification N_{final} whose factual statements are grounded in the CWS.

C. World-State Compiler (WSC)

WSC deterministically compiles $\{\mathcal{F}_k^t\}$ into a globally consistent state $\mathcal{S}^t$ (CWS) that serves as the *single source of truth* for downstream reasoning. WSC is designed to (i) eliminate duplicates, (ii) resolve conflicts under uncertainty, and (iii) preserve provenance for auditing.

1) Cross-sensor normalization and association: Geometry-bearing observations (e.g., LiDAR / radar / BEV agents) are mapped into a unified ego-centric BEV frame. For a 3D point or object center $\mathbf{p}_k$ in sensor frame k,

$$\mathbf{p}_{\text{ego}} = \mathbf{R}_{\text{ego}\leftarrow k}\,\mathbf{p}_k + \mathbf{t}_{\text{ego}\leftarrow k}. \tag{1}$$

If a covariance is provided, WSC propagates it via the Jacobian of the transformation.

Camera detections remain in the image plane and are linked to BEV entities by projecting 3D seeds into the camera view. Let $\mathbf{p}_{\text{ego}} = [x, y, z]^\top$ be a seed corner in the ego frame, and let $\mathbf{K}_{\text{cam}}$ denote the camera intrinsic matrix:

$$\tilde{\mathbf{p}}_{\text{cam}} = \mathbf{R}_{\text{cam}\leftarrow\text{ego}}\mathbf{p}_{\text{ego}} + \mathbf{t}_{\text{cam}\leftarrow\text{ego}}, \quad \mathbf{u} = \pi(\mathbf{K}_{\text{cam}}\tilde{\mathbf{p}}_{\text{cam}}), \tag{2}$$

where $\pi([u', v', w']^\top) = [u'/w',\ v'/w']^\top$. WSC computes the 2D bounding rectangle of the projected 3D box and associates it with camera detections via IoU and class compatibility.

After normalization, WSC performs hierarchical entity association. For a candidate pair (i, j), we define a gated neighborhood:

$$\mathcal{N}(i) = \{j \mid \|\mathbf{p}_i - \mathbf{p}_j\|_2 \le r_g\}, \tag{3}$$

where r_g is a distance gating radius. When covariance is available, a Mahalanobis metric is preferred:

$$d_M(i,j) = (\Delta\mathbf{p})^\top \mathbf{\Sigma}_{ij}^{-1} (\Delta\mathbf{p}), \quad \Delta\mathbf{p} = \mathbf{p}_i - \mathbf{p}_j, \tag{4}$$

where Σ_{ij} denotes the (possibly approximate) joint position covariance used for association in the BEV frame (e.g., $\Sigma_{ij} = P_i + P_j$ after coordinate alignment). A composite cost penalizes class mismatch and size disagreement:

$$C_{ij} = \lambda_p d_M(i,j) + \lambda_c \mathbb{I}[\hat{c}_i \neq \hat{c}_j] + \lambda_s \|\mathbf{s}_i - \mathbf{s}_j\|_2. \tag{5}$$

Here $\hat{c}_i = \arg\max_c p_i(c)$ is the predicted class label of observation i, and $\mathbf{s}_i$ denotes its size vector (e.g., BEV box (l, w) or 3D box (l, w, h) depending on availability) and WSC solves a bipartite assignment under one-to-one constraints to obtain matched clusters. Unmatched observations are preserved with explicit provenance.

2) Reliability-aware fusion and conflict handling: After association, each fused entity may contain multiple sources for the same attribute (e.g., position, velocity, class). WSC resolves attribute conflicts by weighted fusion. For each matched observation $(k, i) \in \mathcal{M}(j)$ and attribute α, we define an unnormalized weight:

$$\tilde{w}_{k,i}^{(\alpha)} = m_k^{(\alpha)} \cdot r_k^{(t)} \cdot q_{k,i} \cdot u_{k,i}^{(\alpha)} \cdot \eta_{k,i}^{(\alpha)}. \tag{6}$$

Here $m_k^{(\alpha)}$ is a modality–attribute prior (e.g., LiDAR for geometry, radar for velocity), $r_k^{(t)}$ is a historical reliability score updated by an exponential moving average, $q_{k,i}$ is the current confidence, $u_{k,i}^{(\alpha)}$ quantifies uncertainty (from $\mathbf{P}$ or a diagonal approximation), and $\eta_{k,i}^{(\alpha)}$ is a cross-source consistency factor. Normalized weights within cluster j are:

$$w_{k,i}^{(\alpha)} = \frac{\tilde{w}_{k,i}^{(\alpha)}}{\sum_{(k',i')\in\mathcal{M}(j)} \tilde{w}_{k',i'}^{(\alpha)}}. \tag{7}$$

Reliability is updated per frame by:

$$r_k^{(t)} = \beta\, r_k^{(t-1)} + (1-\beta)\, s_k^{(t)}, \tag{8}$$

where $s_k^{(t)} \in \{1, 0.5, 0\}$ indicates pass / suspicious / veto under residual gating. For cross-source consistency, we use:

$$\eta_{k,i}^{(\alpha)} = \exp\Big(-d_M\Big(\mathbf{x}_{k,i}^{(\alpha)}, \bar{\mathbf{x}}^{(\alpha)}; \bar{\mathbf{P}}^{(\alpha)}\Big)\Big), \tag{9}$$

with $d_M(\mathbf{x}, \bar{\mathbf{x}}; \bar{\mathbf{P}}) = (\mathbf{x} - \bar{\mathbf{x}})^\top \bar{\mathbf{P}}^{-1} (\mathbf{x} - \bar{\mathbf{x}})$.

a) Continuous attributes: Let $\mathcal{M}(j)$ denote the matched sources for entity j. WSC performs information-weighted fusion:

$$\begin{aligned}
\mathbf{I}_j^{(\alpha)} &= \sum_{(k,i)\in\mathcal{M}(j)} w_{k,i}^{(\alpha)} \left(\mathbf{P}_{k,i}^{(\alpha)}\right)^{-1}, \\
\hat{\mathbf{x}}_j^{(\alpha)} &= \left(\mathbf{I}_j^{(\alpha)}\right)^{-1} \sum_{(k,i)\in\mathcal{M}(j)} w_{k,i}^{(\alpha)} \left(\mathbf{P}_{k,i}^{(\alpha)}\right)^{-1} \mathbf{x}_{k,i}^{(\alpha)}, \\
\hat{\mathbf{P}}_j^{(\alpha)} &= \left(\mathbf{I}_j^{(\alpha)}\right)^{-1}.
\end{aligned} \tag{10}$$

b) Correlated sources: When source lineage indicates potential correlation (e.g., an agent already fuses LiDAR/camera internally), we avoid over-confidence by using Covariance Intersection (CI) [20]:

$$\begin{aligned}
(\mathbf{P}_{\text{CI}})^{-1} &= \omega\, \mathbf{P}_a^{-1} + (1-\omega)\, \mathbf{P}_b^{-1}, \\
\mathbf{x}_{\text{CI}} &= \mathbf{P}_{\text{CI}} \left(\omega\, \mathbf{P}_a^{-1}\mathbf{x}_a + (1-\omega)\, \mathbf{P}_b^{-1}\mathbf{x}_b\right),
\end{aligned} \tag{11}$$

where $\omega \in [0, 1]$ is chosen by a lightweight heuristic (defaulting to balanced fusion and optionally biased toward the lower-residual estimate).

c) Categorical attributes: For discrete attributes (e.g., class), we use weighted voting:

$$p(c \mid j) \propto \sum_{(k,i)\in\mathcal{M}(j)} w_{k,i}^{(\text{cls})}\, p_{k,i}(c), \quad \hat{c}_j = \arg\max_c p(c \mid j). \tag{12}$$

We also attach an ambiguity flag when evidence is weak or inconsistent:

$$\begin{aligned}
p_j^\star &= \max_c p(c \mid j), \\
d_j^\star &= \max_{(k,i)\in\mathcal{M}(j)} d_M(\mathbf{x}_{k,i}, \hat{\mathbf{x}}_j; \hat{\mathbf{P}}_j), \\
\texttt{ambiguous}(j) &= \mathbb{I}\big[p_j^\star < \tau_{\text{cls}} \ \vee \ d_j^\star > \tau_d\big].
\end{aligned} \tag{13}$$

3) CWS generation: Finally, WSC exports the fused entity set as the CWS $\mathcal{S}^t$ in a deterministic schema. Each entity record includes: a global entity ID, fused BEV state (with uncertainty), class label and confidence, sensor provenance and lineage, and ambiguity flags. This design enables transparent tracing from raw evidence to the coordinated world model.

D. Grounding-first, Verifier-aware Driving QA

Given $(Q, \mathcal{S}^t)$ and an optional auxiliary synopsis A, the reasoning engine follows a *grounding-first* principle: $\mathcal{S}^t$ **is binding,** A **is non-binding**. All numeric attributes, entity existence claims, and relations must be supported by entities and fields in $\mathcal{S}^t$.

Recent verification-oriented prompting reduces hallucination by explicitly checking draft answers (e.g., CoVe) [21], yet self-verification can still fail in systematic ways [22]. CoherentDrive therefore combines (i) a structured, conflict-resolved substrate (CWS) with (ii) a verifier loop that rejects ungrounded claims and false premises.

1) Four-stage workflow: We implement a four-stage workflow: (1) **CWS parsing**: extract query-relevant entities/fields from $\mathcal{S}^t$; (2) **risk focusing**: identify right-of-way, proximity, dynamic risks from grounded fields; (3) **draft reasoning**: produce an initial structured decision D and justification N; (4) **introspective verification**: iteratively verify and revise until all claims are grounded or a premise rejection is triggered.

2) Verifier checks: For each atomic claim, the verifier applies three structured checks with deterministic criteria: (i) *entity validity*: referenced IDs exist in $\mathcal{S}^t$; (ii) *field consistency*: cited fields match $\mathcal{S}^t$ (with tolerance for numeric ranges); (iii) *relation support*: relational statements (e.g., "vehicle A is ahead of ego") are derivable from BEV fields. If a query contains a false premise (e.g., refers to a stop sign that does not exist in $\mathcal{S}^t$), the system outputs a premise-rejection response rather than hallucinating supporting evidence. Algorithm 1 summarizes the end-to-end CoherentDrive inference, including WSC compilation and the verifier-aware grounding-first reasoning loop. Here $T_{\text{parse}}, T_{\text{risk}}, T_{\text{reason}}, T_{\text{verify}}, T_{\text{refuse}}, T_{\text{revise}}$ are fixed prompt templates enforcing grounding-first (cite IDs / refuse unsupported premises).

Algorithm 1 CoherentDrive inference: WSC compilation and verifier-aware grounding-first reasoning.

Require: Agent facts $\{F_k^t\}_{k=1}^K$, query Q (optional), synopsis A (optional), max rounds $R_{\max}$
Ensure: D_{final} (JSON), N_{final} (grounded text)
$S \leftarrow \text{WSC}(\{\mathcal{F}_k^t\}_{k=1}^K)$ {S: CWS (IDs/provenance/flags)}
$M \leftarrow \text{LLM}(T_{\text{parse}}, S)$
$L_{\text{risk}} \leftarrow \text{LLM}(T_{\text{risk}}, M)$
$(D, N) \leftarrow \text{LLM}(T_{\text{reason}}, Q, M, L_{\text{risk}}, A)$ {grounding-first}
for $\tau = 1$ **to** $R_{\max}$ **do**
  $V \leftarrow \text{LLM}(T_{\text{verify}}, S, D, N)$ {V: verdict + feedback}
  if V.verdict = `Consistent` **then**
    break
  end if
  if V.verdict = `UnsupportedPremise` **then**
    $(D, N) \leftarrow \text{LLM}(T_{\text{refuse}}, S, Q, V)$; **break**
  end if
  $(D, N) \leftarrow \text{LLM}(T_{\text{revise}}, S, V, D, N)$
end for
$(D_{\text{final}}, N_{\text{final}}) \leftarrow \text{PostProcess}(S, D, N)$
return $(D_{\text{final}}, N_{\text{final}})$

III. Experiments

Our experiments are designed to answer three research questions about interface design for zero-shot driving reasoning. **RQ1** asks whether a coherent, compressed, and conflict-resolved world-state interface (CWS) can simultaneously improve correctness, reduce hallucinations, and lower inference latency (Table I). **RQ2** tests whether these gains come from structuring and conflict resolution rather than merely shorter prompts via length-controlled baselines and ablations (Table II), while **RQ3** stress-tests robustness under cross-modal conflicts and false-premise queries through targeted qualitative cases (Table III).

A. Experimental Setup

Benchmarks. We evaluate on **nuScenes-QA** [10] and a template-aligned **Waymo-QA** split constructed from the Waymo Open Dataset [2] following the same question types and supervision formats as nuScenes-QA. Both benchmarks probe open-world driving question answering and decision reasoning under complex multi-agent traffic scenes.

Methods compared. We include (A) **VLM baselines**: BLIP-2 [8], LLaVA [9], Qwen2-VL [23], and a BEV-aware VLM pipeline (Talk2BEV-style [24]); (B) **agentic driving baselines** including DriveGPT4-style [25] and PlanAgent-style [26]; and (C) **our variants**: *Raw-Logs* (directly concatenating uncoordinated perception logs), *w/o Verifier* (removing the verifier loop, i.e., single-pass grounding-first generation), *w/o WSC* (removing conflict-aware world-state compilation and using a naive union pseudo-state), and *Full* (WSC + CWS + verifier-aware grounding-first protocol).

Metrics. For end-to-end open-ended QA and decision reasoning, we report: *QA Exact Match* (EM), *Answer F1*, *Decision Accuracy* (DA), and *Risk Identification F1* (RI F1) for task-level correctness and safety. For reliability and hallucination, we use *Fact Consistency Pass Rate* (FCPR; higher is better) and *Hallucinated Entity Mention Rate* (HER; lower is better), which measure whether generated answers/decisions are supported by the reference structured scene state and whether the model mentions non-existent entities, respectively. For interface-efficiency analyses, we additionally report *Hallucination Rate* (HR; lower is better), defined as the percentage of answers containing at least one unsupported entity/attribute claim as well as *Average Prompt Length* (APL; in tokens), and *Average Inference Time* (AIT; seconds per query).

Zero-shot protocol. Unless otherwise stated, all text-based methods are evaluated in a zero-shot setting with the same decoding hyperparameters (temperature, max output length). For baselines not fully compatible with our I/O constraints, we implement protocol-aligned reproductions and denote them with the suffix "-style". Table II further reports multi-backbone results with four LLMs to study robustness across model families. In all experiments, we set the optional synopsis $A = \emptyset$ to focus on the effect of the coherent world-state interface.

B. End-to-End Results on nuScenes-QA and Waymo-QA

Table I summarizes end-to-end QA and decision reasoning on nuScenes-QA and Waymo-QA. Overall, **CoherentDrive (Full)** (C4) achieves the best correctness–reliability trade-off across both datasets. On **nuScenes-QA**, C4 reaches **64.3** EM **/ 77.5** F1 and **76.2** DA **/ 66.4** RI F1, while substantially improving reliability (**93.2** FCPR) and suppressing hallucinations (**7.4** HER). Compared to directly prompting uncoordinated raw logs (C1), C4 improves task performance by roughly **+15 pp** on EM/DA and reduces HER by more than **34 pp**, indicating that a coherent, conflict-resolved interface is crucial for zero-shot driving reasoning.

The same trend holds under domain shift on **Waymo-QA**: C4 achieves **61.3** EM **/ 75.2** F1 with **91.4** FCPR and only **8.6** HER, showing strong transfer with only a small degradation (about **3–4 pp** on major task metrics) from nuScenes-QA. Among baselines, Talk2BEV-style is the strongest VLM

TABLE I

END-TO-END PERFORMANCE (%) ON NUSCENES-QA AND WAYMO-QA. HIGHER IS BETTER FOR ALL METRICS EXCEPT HER (LOWER IS BETTER). BEST RESULTS ARE IN **BOLD**.

ID	Method	nuScenes-QA						Waymo-QA					
		EM	F1	DA	RI F1	FCPR	HER	EM	F1	DA	RI F1	FCPR	HER
A1	BLIP-2 [8]	42.3	60.8	50.4	38.5	58.2	36.7	39.2	57.6	46.4	35.7	55.3	38.9
A2	LLaVA [9]	46.1	63.4	52.1	41.2	61.0	34.6	42.1	60.2	48.5	38.4	57.9	36.8
A3	Qwen2-VL [23]	47.8	64.9	54.3	42.0	62.8	33.9	44.3	61.9	50.1	39.2	59.0	35.6
A4	Talk2BEV-style [24]	56.4	70.6	60.8	54.7	74.3	20.8	53.2	67.1	56.3	50.7	71.3	22.7
B1	DriveGPT4-style [25]	52.6	67.3	69.4	58.6	76.1	21.5	49.4	65.2	66.5	55.4	73.2	23.4
B2	PlanAgent-style [26]	55.7	70.2	72.6	63.5	82.4	46.9	52.5	67.4	70.2	60.6	79.4	18.6
C1	CoherentDrive (Raw-Logs)	49.1	66.2	60.3	50.6	55.1	41.8	46.1	63.1	56.0	47.2	52.4	43.7
C2	CoherentDrive (w/o Verifier)	57.3	72.4	68.2	57.4	79.3	17.6	54.2	69.3	65.4	53.5	76.1	19.5
C3	CoherentDrive (w/o WSC)	58.1	73.1	70.4	59.2	81.2	15.9	55.1	70.1	67.2	55.2	78.3	17.9
C4	CoherentDrive (Full)	**64.3**	**77.5**	**76.2**	**66.4**	**93.2**	**7.4**	**61.3**	**75.2**	**73.1**	**62.4**	**91.4**	**8.6**

pipeline, yet it remains notably less reliable than C4 (e.g., much lower FCPR and higher HER). Agentic baselines can yield competitive DA in some cases, but still show weaker factual grounding than C4 as reflected by worse HER/FCPR.

Ablations further confirm that **both** conflict-aware world-state compilation and verifier-aware prompting contribute: removing either the verifier (C2) or WSC (C3) consistently increases HER and decreases FCPR on both datasets. These results collectively support our central claim that *input structuring and compression (via coherent world states)* directly impact correctness, hallucination, and reliability in zero-shot driving QA.

C. Interface Analysis with Multi-Backbone LLMs and Length-Controlled Baselines

We next isolate the effect of *input structure and compression* by comparing raw-log prompting versus Coherent World State (CWS) prompting under a unified *single-pass* QA protocol (verifier disabled for all methods in this table) to isolate the effect of the evidence interface. We report correctness (Accuracy/F1), hallucination (HR), and efficiency (APL/AIT). To address the concern that "shorter prompts trivially help", we introduce two length-controlled baselines for Qwen-14B: (i) **Raw-Logs Truncated**, which truncates the raw prompt to match the CWS token budget, and (ii) **Extractive Condensed Logs**, a deterministic, non-generative compression that de-duplicates and retains only top-ranked nearby agents and core fields within the same budget (thus introducing no new facts).

Table II confirms that structured, conflict-resolved CWS prompting consistently improves zero-shot QA across four LLM backbones, while simultaneously reducing prompt length from ∼1000 to ∼260 tokens and lowering inference latency. Importantly, the length-controlled comparisons show that naive truncation is insufficient: even under the same token budget, truncated raw logs remain unreliable (high HR), whereas extractive condensation improves but still underperforms CWS, indicating that the gains come from *structure + conflict resolution*, not merely fewer tokens.

D. Token Budget Sweep: Quality–Efficiency Pareto

To characterize the trade-off between quality and efficiency, we sweep the evidence token budget and plot accuracy–latency Pareto curves. For CWS, we control the budget by (i) pruning to top-K relevant entities (distance- and query-aware) and (ii) progressively dropping non-essential fields (e.g., fine-grained provenance) under tight budgets. We compare against budget-matched Raw-Logs Truncated and Extractive Condensed Logs.

TABLE II

CONTROLLED INTERFACE STUDY ON A NUSCENES-QA SUBSET (ZERO-SHOT). APL/AIT QUANTIFY EFFICIENCY.

Backbone	Evidence interface	Acc.	F1	HR	APL	AIT
LLaMA2-13B	Raw logs	45	50	30	∼1000	6.5
LLaMA2-13B	CWS (ours, single-pass)	65	70	5	∼260	4.0
Qwen-14B	Raw logs	50	55	25	∼1000	7.0
Qwen-14B	Raw-Logs Truncated (∼260)	47	52	24	∼260	4.5
Qwen-14B	Extractive Condensed (∼260)	62	71	13	∼260	4.6
Qwen-14B	CWS (ours, single-pass)	75	80	5	∼260	4.5
GLM-4-9B	Raw logs	30	35	35	∼1000	4.0
GLM-4-9B	CWS (ours, single-pass)	55	60	10	∼260	2.5
DeepSeek-14B	Raw logs	55	60	20	∼1000	6.0
DeepSeek-14B	CWS (ours, single-pass)	78	85	5	∼260	3.5

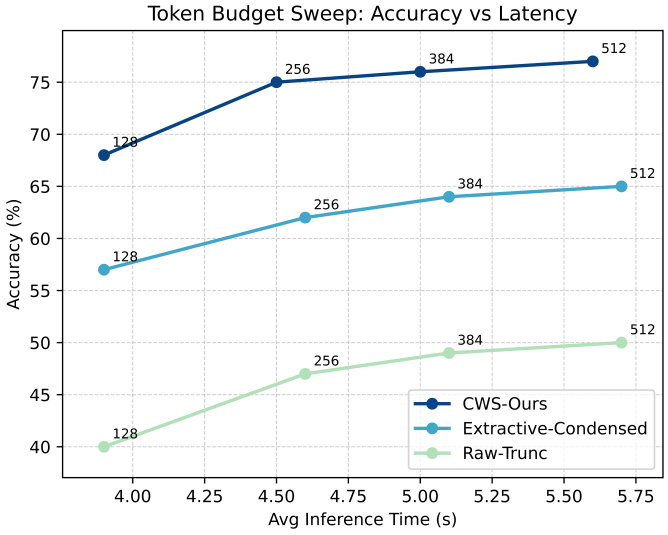
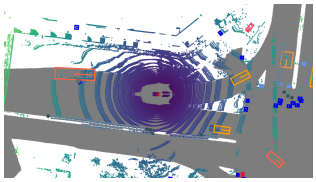
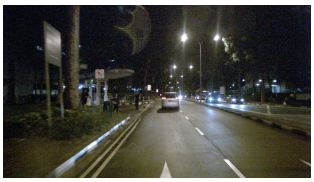
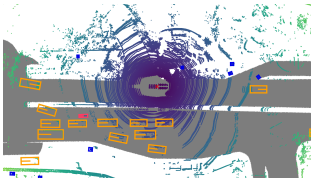


Fig. 2. Token budget sweep on nuScenes-QA (Qwen-14B): accuracy vs. latency. Each marker is labeled by the evidence token budget.

TABLE III

QUALITATIVE CASE STUDIES IN A TOP-DOWN "CASE CARD" STYLE: CASE ID + SCENARIO SUMMARY → QUERIES → BASELINE OUTPUTS → OURS. EACH SCENE INCLUDES A CAMERA VIEW (LEFT) AND A BEV LIDAR VIEW (RIGHT).

Field	Case A: Cross-modal conflict + partial misses	Case B: False-premise query (non-existent bicycle)
Scene	 *Camera*  *BEV-LiDAR*	 *Camera*  *BEV-LiDAR*
Queries	**Q1:** What is the type of the closest leading vehicle ahead (sedan/SUV/van/MPV)? **Q2:** Are there pedestrians near the conflict zone that ego must yield to? What action is recommended?	**Q1:** What are the primary risks in this nighttime scene, and what action should ego take? **Q2:** Is there a bicycle overtaking on the right? If yes, what should ego do?
Baseline	**VLM (Talk2BEV-style):** Q1: "Closest lead vehicle is an SUV." Q2: "No pedestrian detected; proceed cautiously." **Agent (PlanAgent-style):** Q1: "Lead vehicle: SUV (uncertain)." Q2: "Maintain lane and continue (slow down if needed)." (Reason: *conflict* in object attributes and *partial misses* contaminate the evidence.)	**VLM (Qwen2-VL):** Q1: "Low visibility; slow down and keep lane." Q2: "Yes, a bicycle is overtaking; give way." **Agent (DriveGPT4-style):** Q1: "Nighttime; reduce speed and stay centered." Q2: "Bicycle on the right; slow down and yield." (Reason: a *false premise* leads to a hallucinated bicycle and unnecessary reactions.)
Ours	**CoherentDrive (Full):** Q1: "Closest lead vehicle is a van/MPV (conflict-resolved, ID-grounded)." Q2: "Pedestrians present near the conflict zone (ID-grounded). Action: slow down and yield/stop." (Why it works: WSC resolves conflicts into a coherent state; verifier enforces ID-grounded answers.)	**CoherentDrive (Full):** Q1: "Risks: low visibility + VRUs + limited stopping margin (ID-grounded). Action: slow down and keep lane, be ready to brake." Q2: "Unsupported premise: no bicycle entity exists. Action: do not react to a non-existent bicycle; keep observing." (Why it works: verifier checks the premise against the coherent world state and rejects unsupported entities.)

E. Qualitative Analysis

We qualitatively analyze two representative failure modes that frequently degrade zero-shot driving reasoning: **(i) cross-modal conflicts and partial misses** that contaminate the evidence interface, and **(ii) false-premise queries** that induce premise-driven hallucinations. Table III summarizes two cases (A and B) with identical questions and representative baseline outputs.

Case A: conflict-sensitive attribute + safety-critical completeness. This case contains (1) a leading vehicle whose fine-grained type is reported inconsistently across sources (e.g., SUV vs. van/MPV), and (2) pedestrians near the crosswalk/conflict zone that require conservative behavior. When prompted with raw logs, the LLM tends to hedge or remain uncertain, because the prompt itself is internally inconsistent. BEV/VLM baselines may commit to a single answer, but the commitment is difficult to audit and may vary across methods. In contrast, CoherentDrive resolves conflicts into a single coherent world state and retains single-source detections for completeness, allowing the final answer to (a) commit to a consistent grounded type and (b) recommend a conservative action based on explicitly supported entities.

Case B: false-premise attack (non-existent bicycle). Here we intentionally query a bicycle overtaking on the right, which is absent in the scene. A typical failure mode is *premise acceptance*: the model answers "yes" and recommends avoidance actions for a non-existent target, which is unsafe in a high-stakes setting. Some structured agents may hedge ("cannot confirm"), yet still condition actions on an unverified premise. CoherentDrive explicitly verifies the premise against the coherent world state and rejects unsupported entities, thereby preventing premise-driven hallucinations from propagating into decisions.

These qualitative patterns align with our quantitative findings: (1) interface cleanliness (conflict-resolved state vs. raw logs) is crucial for correctness and for reducing hallucinations, and (2) verifier-aware prompting is a necessary second line of defense under misleading or adversarial queries. Together, they explain why C4 achieves both higher FCPR and much lower HER in Table I.

IV. CONCLUSION

In this paper, we presented CoherentDrive, a coordination-first grounding framework for hallucination-resistant driving question answering. CoherentDrive treats reliability as an interface design problem: a deterministic World-State Compiler (WSC) compiles heterogeneous and potentially conflicting perception logs into a compact, conflict-resolved Coherent World State (CWS), and a grounding-first, verifier-aware protocol restricts generation to CWS-supported entities, relations, and numeric fields with explicit premise rejection. Extensive evaluations on nuScenes-QA and a template-aligned Waymo-QA split demonstrate that CoherentDrive improves task correctness and safety-related reasoning while substantially reducing hallucinated entity mentions, and controlled studies further verify that the gains stem from structured, conflict-aware world-state prompting rather than merely shorter prompts. We believe coherent world-state interfaces provide a practical foundation for integrating LLM reasoning into safety-critical driving interactions.

Limitations and future work. Our study assumes upstream perception agents provide reasonably calibrated uncertainty and does not yet address closed-loop driving control. Future work includes extending world-state compilation to longer temporal horizons, learning reliability priors from data, and integrating CWS-grounded reasoning with planning and policy learning for end-to-end autonomous driving.

REFERENCES

[1] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 621–11 631.

[2] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *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, pp. 2446–2454.

[3] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[4] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 2, 2023, pp. 1477–1485.

[5] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 2774–2781.

[6] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1090–1099.

[7] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 23 716–23 736, 2022.

[8] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International Conference on Machine Learning*. PMLR, 2023, pp. 19 730–19 742.

[9] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 34 892–34 916, 2023.

[10] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4542–4550.

[11] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving with graph visual question answering," in *European Conference on Computer Vision*. Springer, 2024, pp. 256–274.

[12] K. Renz, L. Chen, E. Arani, and O. Sinavski, "Simlingo: Vision-only closed-loop autonomous driving with language-action alignment," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 11 993–12 003.

[13] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[14] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 1906–1919.

[15] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.

[16] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *11th International Conference on Learning Representations, ICLR 2023*, 2023.

[17] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 8634–8652, 2023.

[18] Y. Tu, R. Hu, and J. Sang, "Ode: Open-set evaluation of hallucinations in multimodal large language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 19 836–19 845.

[19] F. Tang, C. Liu, Z. Xu, M. Hu, Z. Huang, H. Xue, Z. Chen, Z. Peng, Z. Yang, S. Zhou *et al.*, "Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 26 147–26 159.

[20] S. J. Julier and J. K. Uhlmann, "A non-divergent estimation algorithm in the presence of unknown correlations," in *Proceedings of the 1997 American Control Conference (Cat. No. 97CH36041)*, vol. 4. IEEE, 1997, pp. 2369–2373.

[21] S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston, "Chain-of-verification reduces hallucination in large language models," in *Findings of the association for computational linguistics: ACL 2024*, 2024, pp. 3563–3578.

[22] K. Stechly, K. Valmeekam, and S. Kambhampati, "On the self-verification limitations of large language models on reasoning and planning tasks," in *The Thirteenth International Conference on Learning Representations*.

[23] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[24] T. Choudhary, V. Dewangan, S. Chandhok, S. Priyadarshan, A. Jain, A. K. Singh, S. Srivastava, K. M. Jatavallabhula, and K. M. Krishna, "Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 16 345–16 352.

[25] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, and H. Zhao, "Drivegpt4: Interpretable end-to-end autonomous driving via large language model," *IEEE Robotics and Automation Letters*, 2024.

[26] Y. Zheng, Z. Xing, Q. Zhang, B. Jin, P. Li, Y. Zheng, Z. Xia, K. Zhan, X. Lang, Y. Chen *et al.*, "Planagent: A multi-modal large language agent for closed-loop vehicle motion planning," *arXiv preprint arXiv:2406.01587*, 2024.