

OmniReason: A Temporal-Guided Vision-Language-Action Framework for Autonomous Driving

Pei Liu^{1*}, Qingtian Ning^{2*}, Xinyan Lu², Haipeng Liu², Weiliang Ma², Dangen She², Peng Jia²,
Xianpeng Lang², Jun Ma^{1,3†}

¹The Hong Kong University of Science and Technology (Guangzhou)

²Li Auto Inc.

³The Hong Kong University of Science and Technology

Abstract

Recent advances in vision-language models (VLMs) have demonstrated impressive spatial reasoning capabilities for autonomous driving, yet existing methods predominantly focus on static scene understanding while neglecting the essential temporal dimension of real-world driving scenarios. To address this critical limitation, we propose the OmniReason framework, which establishes robust spatiotemporal reasoning by jointly modeling dynamic 3D environments and their underlying decision-making processes. Our work makes two fundamental advances: (1) We introduce OmniReason-Data, two large-scale vision-language-action (VLA) datasets with dense spatiotemporal annotations and natural language explanations, generated through a novel hallucination-mitigated auto-labeling pipeline that ensures both physical plausibility and temporal coherence; (2) We develop the OmniReason-Agent architecture, which integrates a sparse temporal memory module for persistent scene context modeling and an explanation generator that produces human-interpretable decision rationales, facilitated by our spatiotemporal knowledge distillation approach that effectively captures spatiotemporal causal reasoning patterns. Comprehensive experiments demonstrate state-of-the-art performance, where OmniReason-Agent achieves significant improvements in both open-loop planning tasks and visual question answering (VQA) benchmarks, while establishing new capabilities for interpretable, temporally-aware autonomous vehicles operating in complex, dynamic environments.

Introduction

End-to-end (E2E) learning has rapidly become a foundational paradigm for autonomous driving, facilitating the joint optimization of perception, prediction, and planning within a unified framework (Zhou et al. 2024; Chen et al. 2024c). Leveraging large-scale driving datasets, these models are capable of learning driving policies directly from raw sensor data, delivering impressive performance across a spectrum of real-world scenarios. However, despite these achievements, current E2E approaches encounter persistent challenges: they often struggle to generalize to rare long-tail events, exhibit insufficient comprehension of high-level

scene semantics, and lack the adaptive, interpretable reasoning skills required in open-world environments (Chen et al. 2024b).

Simultaneously, the advent of large language models (LLMs) and vision-language models (VLMs) has underscored their remarkable abilities in in-context learning, commonsense reasoning, and generalization beyond the training distribution. These emergent capabilities present a tantalizing opportunity for enhancing the intelligence and robustness of autonomous driving systems, especially when confronted with the demanding complexities of real-world, safety-critical deployment. (Liu et al. 2024b; Zhou and Knoll 2024). However, directly applying existing VLMs to autonomous driving presents significant challenges. Most VLMs are primarily optimized for static 2D vision-language tasks, which limits their capacity for spatial reasoning and comprehensive scene understanding in rich, dynamic 3D driving environments (Radford et al. 2021; Zhai et al. 2023). More critically, the absence of explicit temporal modeling mechanisms prevents these models from effectively reasoning about interactions, motion, and causal relationships as they unfold over time (Zhou et al. 2025). In addition, their tendency to produce hallucinated or unreliable descriptions undermines the overall trustworthiness needed for high-stakes autonomous applications (Zhang et al. 2025; Xu, Jain, and Kankanhalli 2024). Consequently, a major technical challenge emerges: **How can we effectively leverage the generalization and reasoning capabilities of large-scale VLMs to develop Vision-Language-Action (VLA) frameworks that generate reliable, temporally coherent, and interpretable driving actions in rapidly changing real-world conditions?**

We introduce OmniReason-Data, which comprises two large-scale, human-like VLA datasets: OmniReason-nuScenes and OmniReason-Bench2Drive that provide rich, temporally dense annotations integrating human prior knowledge, causal reasoning chains, and scene-aware descriptions across diverse and challenging driving scenarios. Leveraging multimodal supervision and natural language rationales, OmniReason advances the realism, interpretability, and coverage of autonomous driving benchmarks, laying a strong foundation for scene understanding and action generation research.

*These authors contributed equally.

†Corresponding author.

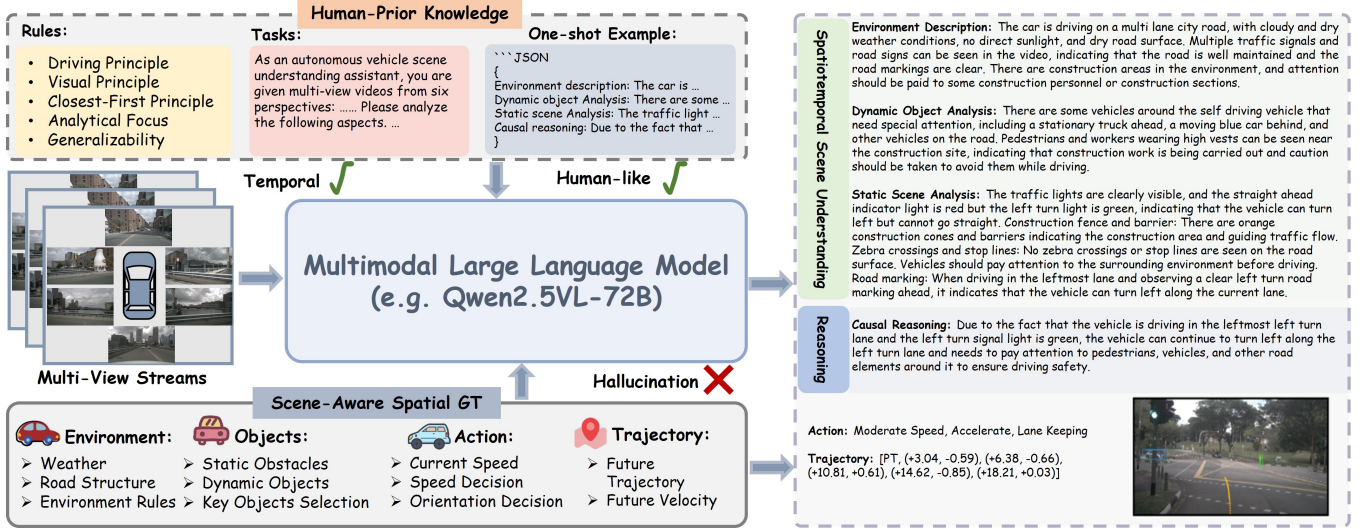


Figure 1: Annotation pipeline. We adopt a rule-based QA labeling pipeline, where the ground truth annotations in nuScenes and Bench2Drive are used. A critical part of our pipeline is the ground truth elements, which contain environment, objects, action, and trajectory, guaranteeing high data quality at spatiotemporal understanding and reasoning.

Building upon these datasets, we further propose OmniReason-Agent, a novel E2E VLA model that is explicitly designed to leverage the multifaceted supervision present in OmniReason datasets. Crucially, OmniReason-Agent encodes and injects human-like priors and spatiotemporal causal reasoning throughout the entire perception, planning, and action pipeline via a knowledge distillation framework. By distilling structured rationales, scene analyses, and expert decision patterns from the comprehensive OmniReason datasets into the agent, our model can internalize human-centric knowledge and causal inference, resulting in more interpretable, reliable, and context-aware autonomous driving behavior. Extensive experiments and ablation studies demonstrate that this approach achieves state-of-the-art results across metrics of safety, comfort, and interpretability, substantially narrowing the gap between autonomous driving systems and experienced human drivers in both decision quality and explanation capability. Our contributions are summarized as follows:

- We introduce OmniReason-nuScenes and OmniReason-Bench2Drive, two comprehensive VLA datasets that emphasize causal reasoning grounded in both spatial and temporal context across diverse and complex driving scenarios. By integrating human prior knowledge, step-wise reasoning over dynamic scenes, and detailed context-aware descriptions, OmniReason establishes a new standard for interpretability and realism in autonomous driving research.
- Our novel template-driven annotation framework is built upon structured scene analysis, enabling the automated generation of high-quality and interpretable language-action pairs for diverse driving scenarios. This approach ensures accurate supervision, mitigates hallucination artifacts, and delivers rich multimodal rationales that are

well-suited for both model training and thorough evaluation.

- We design OmniReason-Agent, an E2E VLA architecture that infuses temporal 3D knowledge and causal reasoning via a knowledge distillation framework. This enables the agent to internalize expert decision patterns and natural language rationales, resulting in context-aware and highly interpretable autonomous driving behavior.
- Extensive experiments and ablation studies on open-loop planning and driving visual question answering (VQA) benchmarks demonstrate that OmniReason-Data and OmniReason-Agent achieve state-of-the-art results in safety, comfort, and explanation quality. Our approach significantly narrows the gap between autonomous vehicles and human-driven vehicles, enhancing reliability and understanding in real-world autonomous driving scenarios.

Related Work

Language-enhanced Dataset for Autonomous Driving

Current dataset evolution demonstrates a clear trajectory from single-task annotation (e.g., object recognition) toward comprehensive scene understanding, incorporating multi-sensor streams and addressing complex challenges like motion prediction and E2E driving. Recent breakthroughs in LLMs have spurred the creation of vision-language driving datasets aiming to capture environmental semantics and action rationales (Li et al. 2023; Liu et al. 2024b). BDD-X (Kim et al. 2018) pioneers free-form captions explaining ego-vehicle motivations, while BDDOIA (Xu et al. 2020) provides closed-form explanations for object-induced responses. DRAMA (Malla et al. 2023) extends this with risk-

Dataset	Vision Data		Language Data		Action Data Type	Reason	Weat	Road	Trfc.	Trfc.	Trfc.	Sur.	Obj.
	M.V.	Temporal	VQA	Auto.		-ing	-her	Type	Flow	Light	Sign	Obj.	Pos.
Talk2Car (Deruyttere et al. 2019)	✗	✗	✗	✗	Traj.	✗	✗	✗	✗	✗	✗	✓	✗
T2C-Traj (Deruyttere et al. 2022)	✗	✗	✗	✗	Traj., Command	✗	✗	✗	✗	✗	✗	✓	✗
DriveLM-nuScenes (Sima et al. 2024)	✗	✗	✓	✗	Traj.	✓	✗	✗	✗	✗	✗	✓	✗
DriveLM-CARLA (Sima et al. 2024)	✓	✗	✓	✓	Traj.	✓	✗	✗	✗	✗	✗	✓	✗
DRAMA (Malla et al. 2023)	✗	✓	✓	✓	Command	✓	✗	✓	✗	✓	✓	✓	✗
Rank2Tell (Sachdeva et al. 2024)	✗	✓	✗	✓	Command	✓	✗	✗	✗	✓	✗	✓	✓
Reason2Drive (Nie et al. 2024)	✗	✗	✓	✓	Command	✓	✗	✗	✗	✗	✗	✓	✗
Omnidrive (Wang et al. 2025)	✓	✗	✓	✓	Traj.	✓	✗	✓	✗	✓	✓	✓	✓
CoVLA (Arai et al. 2025)	✗	✗	✗	✓	Traj.	✗	✓	✗	✗	✓	✗	✓	✗
OmniReason-nuScenes (Ours)	✓	✓	✓	✓	Traj., Command	✓	✓	✓	✓	✓	✓	✓	✓
OmniReason-Bench2Drive (Ours)	✓	✓	✓	✓	Traj., Command	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison of E2E autonomous driving datasets with language and action data. M.V.: Multi-View images; Auto.: Auto-labeling; Trfc.: Traffic; Sur.: Surrounding; Obj.: Object; Pos.: Position; Traj.: Trajectory.

aware annotations and safety-critical commands. OpenDV-2K (Yang et al. 2024) scales up through automatically generated text-driven action pairs from public videos. However, these resources primarily offer high-level discrete commands without precise trajectory specifications essential for detailed maneuver representation.

Emerging datasets like HAD (Kim et al. 2019), Talk2Car (Deruyttere et al. 2019), Talk2Car-Trajectory (Deruyttere et al. 2022), DriveLM (Sima et al. 2024), Omnidrive (Wang et al. 2025), and CoVLA (Arai et al. 2025) attempt to address this gap by incorporating trajectory-level action annotations. Despite these efforts, their limited diversity in temporal and spatial driving scenarios restricts the ability to fully evaluate and train robust VLA models with comprehensive causal reasoning capabilities. A thorough comparison between existing prompt-based driving datasets and ours is summarized in Table 1.

MLLMs Grounded Autonomous Driving

Recent research in autonomous driving has witnessed growing interest in leveraging multimodal large language models (MLLMs) through diverse architectural paradigms. Prompt-based approaches such as GPT-Driver (Mao et al. 2023a) and LLM-Driver (Chen et al. 2024a) operationalize scene understanding by encoding perceptual states into textual prompts for high-level planning, while E2E frameworks like DriveGPT4 (Xu et al. 2024) directly tokenize raw sensor inputs to predict control signals. Hybrid architectures exemplified by LMDrive (Shao et al. 2024), DriveMLM (Wang et al. 2023b), and CarLLaVA (Renz et al. 2024) employ cascaded designs that bridge visual encoders with specialized control decoders through intermediate feature fusion. Despite these methodological advancements, existing systems exhibit inherent limitations in modeling causal relationships between dynamic environmental stimuli and vehicular responses, a critical capability for explainable decision-making in complex scenarios. Our work addresses this gap by integrating vision-language models with causal reasoning question-answering mechanisms, establishing an interpretable framework that reasons through spatiotemporal dependencies rather than surface-level correlations. This

approach uniquely enables zero-shot E2E planning capabilities, advancing beyond conventional stimulus-response mappings demonstrated in prior studies (Han et al. 2025; Mao et al. 2023b; Marcu et al. 2023; Pan et al. 2024; Seff et al. 2023; Yuan et al. 2024) through explicit grounding in causal-temporal reasoning principles.

OmniReason-Data

We propose OmniReason-nuScenes and OmniReason-Bench2Drive, two datasets founded on spatiotemporal causal reasoning within the nuScenes and Bench2Drive frameworks. These datasets comprise high-quality VQA pairs that encompass static and dynamic captions, as well as causal reasoning and action within a four-dimensional domain. OmniReason leverages human-prior knowledge, scene-aware spatial ground truth, multi-view streams, and MLLMs for spatiotemporal understanding and causal reasoning, as illustrated in Fig. 1.

Data Annotation Pipeline

As illustrated in Fig. 1, our data annotation pipeline is designed to construct OmniReason through a unified, scene-aware VQA paradigm that tightly integrates human prior knowledge, task decomposition, and large language model reasoning. The process starts with synchronized multi-view video streams, comprehensive object-level spatial annotations, and ego-vehicle state captured from globally-sourced benchmarks (nuScenes (Caesar et al. 2020) and Bench2Drive (Jia et al. 2024)). These raw data sources are parsed into an ego-centric temporal database that serves as the foundation for downstream annotation. Each frame is organized around the driving context and encodes granular information such as environment descriptors, static and dynamic objects, action, and trajectory, mirroring the rich spatial and temporal signals observed by human drivers.

On top of this structured database, we apply meticulously crafted, rule-based and principle-based templates, reflecting human prior driving knowledge and explicit reasoning focuses, to guide the generation of language-action pairs. Our framework encompasses scenario-level prompts that instruct the model to analyze the driving context from

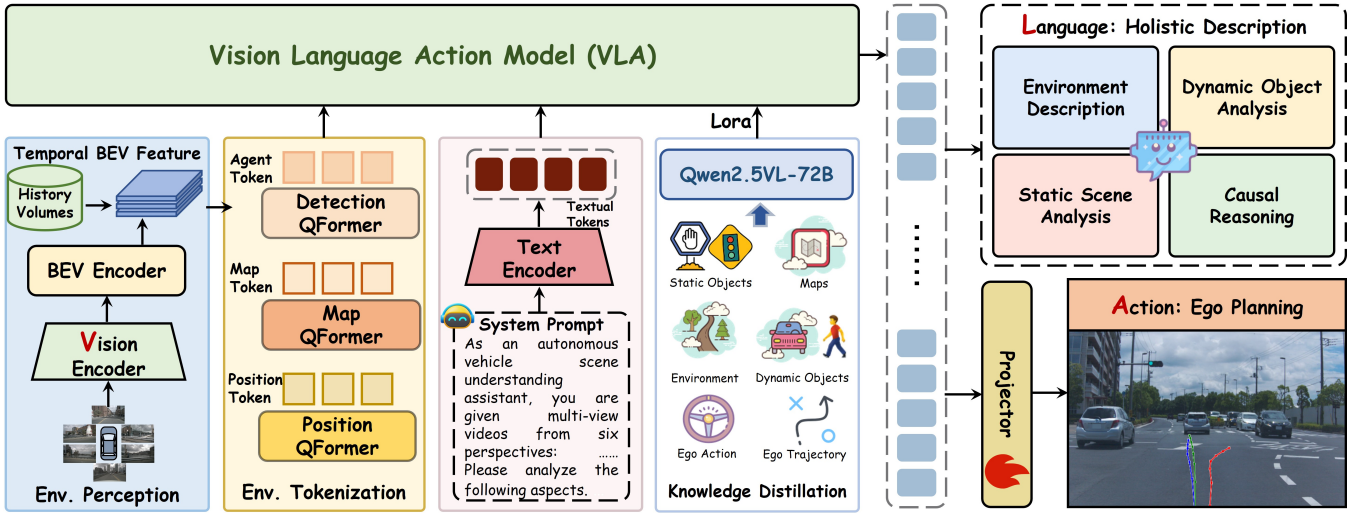


Figure 2: OmniReason harnesses the power of open-source pre-trained language foundation models to generate context-aware driving actions by integrating 3D environmental perception, real-time ego vehicle states, and natural language driver commands. This multimodal integration enables the system to achieve state-of-the-art performance across both open-loop trajectory planning and complex vision-language driving tasks, effectively demonstrating its dual capabilities in precise motion generation and sophisticated scene understanding. The model’s strong performance highlights its ability to bridge perceptual inputs with actionable driving decisions while maintaining robust reasoning about dynamic traffic scenarios.

multiple perspectives, such as environment understanding, static scene analysis, dynamic object analysis, and action-trajectory planning. The Qwen2.5VL 72B (Bai et al. 2025) multimodal language model then receives these multi-view spatial and temporal inputs, structured prompts, and explicit human-like task requirements. As visualized in Fig. 1, the model is tasked with generating not only accurate scene descriptions but also multi-step causal chains that explicate how agent behavior, environmental context, and regulatory factors interact over time.

Our pipeline systematically integrates spatiotemporal understanding, causal reasoning, and future action to generate comprehensive annotation outputs featuring both fine-grained scene analysis and interpretable chain-of-thought rationales firmly grounded in observational data. The generated samples are structured through carefully designed question templates to produce diverse VQA pairs that are both robust against hallucination and highly informative for autonomous driving model training and evaluation. This approach ensures the resulting dataset captures rich temporal dynamics and causal relationships while maintaining strict alignment with real-world driving scenarios. More details and additional analyses are provided in **Appendix A**.

OmniReason-Agent

We further present OmniReason-Agent, an E2E autonomous driving framework that unifies vision-language understanding with actionable trajectory generation through three coordinated components, as depicted in Fig. 2. The pipeline first encodes multi-view camera inputs into spatiotemporal tokens via a hierarchical vision backbone, where our sparse temporal memory module employs memory-compressed

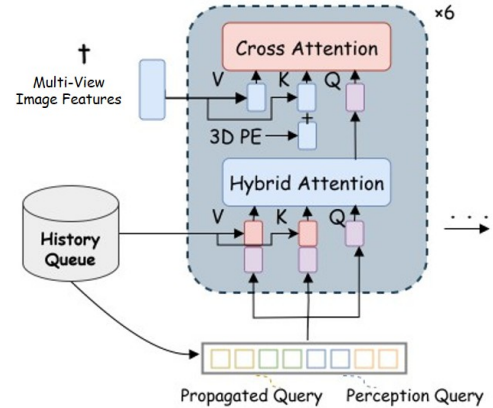


Figure 3: The sparse temporal model employs an iterative architecture where object queries dynamically interact with hybrid queries and multi-view image features for progressive feature refinement. PE denotes Position Embedding.

queries to aggregate long-range visual context while reducing computational complexity. The VLM reasoning core then processes the distilled scene features alongside historical driving states and natural language instructions, executing real-time decision-making through a frozen VLM augmented with lightweight adapters.

Sparse Temporal Memory Module

As illustrated in Fig. 3, the temporal modeling architecture integrates motion compensation, memory-aware feature

propagation, and multi-modal fusion through three synergistic components, inspired by (Wang et al. 2023a). The center of this framework is the motion-aware normalization module, which dynamically adapts object states by encoding spatiotemporal context via the transformation:

$$\mathcal{E}_{t-1}^t = \mathcal{E}_t \cdot \mathcal{E}_{t-1}, \quad (1)$$

$$\mathcal{Q}_p^t = \mathcal{E}_{t-1}^t \cdot \mathcal{Q}_p^{t-1}, \quad (2)$$

where $\mathcal{E}_{t-1}^t$ represents the ego-motion transformation matrix between consecutive frames, and $\mathcal{Q}_p^t$ denotes the temporally aligned 3D object centers from historical observations. This geometric alignment is enhanced through conditional affine transformations that parameterize motion attributes ($v, \Delta t$) into normalization coefficients:

$$\alpha = \xi_1(\mathcal{E}_{t-1}^t, v, \Delta t), \quad (3)$$

$$\beta = \xi_2(\mathcal{E}_{t-1}^t, v, \Delta t), \quad (4)$$

modulating both position encodings and contextual features as:

$$\mathcal{Q}_{pe}^t = \alpha \cdot LN(\psi(\mathcal{Q}_p^t)) + \beta, \quad (5)$$

$$\mathcal{Q}_c^t = \alpha \cdot LN(\mathcal{Q}_c^t) + \beta, \quad (6)$$

where ψ denotes an MLP that projects aligned coordinates into positional embeddings, and LN means layer normalization. The hybrid attention mechanism then establishes temporal coherence by propagating object queries across frames. Perception queries $\mathcal{Q}_c$ merge with propagated queries $\mathcal{Q}_m$ to form a compact interaction space:

$$\mathcal{Q}' = MHA([\mathcal{Q}_m, \mathcal{Q}_c], [\mathcal{Q}_m, \mathcal{Q}_c], [\mathcal{Q}_m, \mathcal{Q}_c]), \quad (7)$$

where MHA means multi-head attention, enabling efficient attention queries while suppressing redundant detections. Cross-modal aggregation subsequently fuses image features $\mathcal{F}_{img}$ with temporal object queries via:

$$\mathcal{Q}' = CA([\mathcal{Q}_m, \mathcal{Q}_c], \mathcal{Q}_{pe}^t + \mathcal{F}_{img}, \mathcal{F}_{img}), \quad (8)$$

where CA stands for cross-attention. The attention mechanism is architecturally agnostic—compatible with global attention from PETR (Liu et al. 2022) or projective attention from DETR3D (Wang et al. 2022). This triad of components achieves temporally consistent perception by (1) compensating ego-motion through geometric alignment, (2) propagating object states via memory-conditioned normalization, and (3) maintaining cross-frame identity through query propagation. The unified design ensures stable object trajectories under complex scenarios involving occlusion and abrupt motion changes.

Training Objectives

The proposed model employs a composite loss function to jointly optimize 3D object detection and structured scene understanding. The detection objective combines categorical recognition and spatial localization through two key terms: the classification loss $\mathcal{L}_{cls}$ for object categories is formulated via Focal Loss, while the regression loss $\mathcal{L}_{reg}$ for 3D bounding box coordinates adopts an L1 formulation encoding center coordinates, dimensions, and orientation angles.

For lane and road structure analysis, the framework applies analogous supervision with lane classification loss $\mathcal{L}_{mcls}$ and geometric regression loss $\mathcal{L}_{mreg}$, each scaled by task-specific coefficients. The total loss of QFormer is:

$$\mathcal{L}_{pc} = \lambda_c \mathcal{L}_{cls} + \lambda_r \mathcal{L}_{reg} + \lambda_{mc} \mathcal{L}_{mcls} + \lambda_{mr} \mathcal{L}_{mreg}, \quad (9)$$

where $\lambda_c, \lambda_r, \lambda_{mc}$, and λ_{mr} represent loss balancing coefficients for detection classification, detection regression, lane classification, and lane regression tasks, respectively.

For the LLM, we leverage the auto-regressive cross-entropy loss $\mathcal{L}_{ce}$. The unified training objective aggregates these components as:

$$\mathcal{L}_{total} = \mathcal{L}_{pc} + \mathcal{L}_{ce}. \quad (10)$$

Experiments

Datasets and Evaluation Metrics

The proposed OmniReason dataset encompasses open-loop planning, closed-loop planning, and VQA tasks. We evaluate OmniReason-Agent on the open-loop planning task using the nuScenes benchmark, measuring planning performance via L2 displacement errors at 1, 2, and 3 seconds, along with the average collision rate (CR) and intersection rate (IR) over the prediction horizon. To further probe the model’s scene understanding and reasoning capabilities, we assess its performance on our driving VQA dataset after instruction tuning. For VQA evaluation, we adopt standard metrics: CIDEr (CI-r), BLEU-1 (BL-1), BLEU-4 (BL-4), METEOR (ME-R), ROUGE-L (RO-L), Precision, and Recall, ensuring a comprehensive analysis of language understanding and multimodal alignment.

Implementation Details

All experiments are conducted using 128 NVIDIA H20 GPUs, each with 96GB of memory. For visual feature extraction, we employ EVA-02-L (Fang et al. 2024) as our vision encoder, which is pretrained through masked image modeling with knowledge distillation from CLIP (Radford et al. 2021). This approach ensures strong alignment between visual features and language representations. LLaVA v1.5 (Liu et al. 2024a) serves as our base model, and we adopt the same dataset composition and training configuration for the 2D pretraining phase. During the fine-tuning stage, we optimize the model using AdamW (Loshchilov and Hutter 2016) with a batch size of 16. We apply differentiated learning rates: 4×10^{-4} for the projector module, while keeping both the visual encoder and large language model at a lower learning rate of 2×10^{-5} to preserve their pretrained knowledge. The entire fine-tuning process utilizes a cosine annealing schedule to maintain training stability and achieve optimal convergence.

Open-Loop Trajectory Planning Task

Table 2 presents a detailed comparison of open-loop planning methods on the nuScenes dataset. The proposed OmniReason-Agent achieves the lowest average L2 distance of 0.34 m, matching the top-performing ORION method and outperforming other baselines such as BEV-Planner with

Method	L2 (m) ↓				CR (%) ↓				IR (%) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.	1s	2s	3s	Avg.
ST-P3 (Hu et al. 2022)	1.59	2.64	3.73	2.65	0.69	3.62	8.39	4.23	2.53	8.17	14.4	8.37
VAD-Base (Jiang et al. 2023)	0.69	1.22	1.83	1.25	0.06	0.68	2.52	1.09	1.02	3.44	7.00	3.82
BEV-Planner (Li et al. 2024b)	0.30	0.52	0.83	0.55	0.10	0.37	1.30	0.59	0.78	3.79	8.22	4.26
DriveVLM† (Tian et al. 2024)	0.18	0.34	0.68	0.40	0.10	0.22	0.45	0.27	-	-	-	-
Senna (Jiang et al. 2024)	0.26	0.42	0.61	0.43	0.05	0.11	0.21	0.12	-	-	-	-
Omni-L (Wang et al. 2025)	1.43	2.34	3.24	2.34	0.23	1.47	4.00	1.90	0.90	2.82	6.16	3.29
Omni-Q	1.15	1.96	2.84	1.98	0.80	3.12	7.46	3.79	1.66	3.86	8.26	4.59
ORION (Fu et al. 2025)	0.17	0.31	0.55	0.34	0.05	0.25	0.80	0.37	-	-	-	-
OmniReason-Agent	0.15	0.31	0.57	0.34	0.04	0.18	0.98	0.40	0.61	2.75	6.19	3.18

Table 2: Comparisons of open-loop planning results on the nuScenes dataset. † indicates models where both ego status and trajectory planning are processed through textual modality using LLMs, while ‡ denotes the exclusion of high-level command inputs during training and inference. All compared approaches maintain fair evaluation standards by avoiding explicit ego status disclosure in their generative planners.

Model	OmniReason-nuScenes					OmniReason-Bench2Drive				
	CI-r	BL-1	BL-4	ME-R	RO-L	CI-r	BL-1	BL-4	ME-R	RO-L
Qwen2.5VL 72B (Bai et al. 2025)	67.14	18.78	3.25	20.75	21.91	87.57	28.03	5.81	67.27	28.10
Qwen2.5VL 32B	59.37	15.88	1.72	17.69	19.13	83.80	30.76	5.05	58.09	25.84
Qwen2.5VL 7B	62.78	19.86	2.96	22.52	22.34	75.65	14.55	2.20	71.03	21.88
Qwen2.5VL 3B	48.59	19.08	1.91	22.51	21.13	67.21	6.39	0.92	77.27	19.07
Qwen2VL 72B (Team 2024)	57.10	25.76	4.46	31.81	26.56	78.60	8.21	1.67	83.38	22.60
InternVL3 14B (Zhu et al. 2025)	70.01	8.82	1.09	74.15	19.18	78.55	9.78	1.71	78.42	22.2
InternVL3 8B	64.64	6.41	0.67	74.02	16.7	70.58	7.33	1.1	76.56	19.42
LLaVA_NEXT 7B (Li et al. 2024a)	53.54	4.71	0.47	76.09	15.63	60.08	5.35	0.63	78.53	17.51
OmniReason-Agent	92.39	51.54	14.45	64.45	32.75	95.46	47.25	18.06	58.78	36.99

Table 3: Comparison of VQA performance conducted on the OmniReason-nuScenes and OmniReason-Bench2Drive datasets. The results show that OmniReason-Agent consistently outperforms previous state-of-the-art open-source MLLMs across most metrics. Notably, all reported indicators are positively oriented, with higher values indicating superior performance.

0.55 m, DriveVLM with 0.40 m, and VAD-Base with 1.25 m. In terms of collision rate, OmniReason-Agent attains an average value of 0.40%, which is slightly higher than Senna’s best result of 0.12% and ORION’s 0.37%, but remains lower than most other baselines. For infraction rate, OmniReason-Agent sets a new state-of-the-art with an average of 3.18%, outperforming all previous methods reporting this metric, including BEV-Planner with 4.26%, VAD-Base with 3.82%, and Omni-Q with 4.59%. Across all evaluated horizons, OmniReason-Agent demonstrates superior or highly competitive performance in both safety and trajectory accuracy, highlighting the effectiveness of its reasoning framework under fair evaluation protocols that prevent explicit ego status leakage.

Driving Visual Question Answering Task

As shown in Table 3, OmniReason-Agent achieves significant performance improvements across all metrics on both OmniReason-nuScenes and OmniReason-Bench2Drive benchmarks compared to state-of-the-art open-source MLLMs. On OmniReason-nuScenes, our method demonstrates particularly strong results with a 37.6% relative improvement in CIDEr, while showing even more dramatic gains of 100.1% and 224.0% in

BLEU-1 and BLEU-4 scores, respectively. The performance on OmniReason-Bench2Drive remains consistently superior, with OmniReason-Agent achieving 9.0%, 53.6%, and 210.8% relative improvements in CIDEr, BLEU-1, and BLEU-4 metrics over the best baseline. Notably, these substantial advances are consistently reflected across both datasets in the ROUGE-L metric as well. These comprehensive results not only validate OmniReason-Agent’s enhanced multimodal reasoning capabilities but also establish new state-of-the-art performance for VQA tasks in autonomous driving scenarios.

Ablation Study

Analysis on Different Language Components. Table 4 presents a detailed evaluation of OmniReason-Agent’s performance across different language components in the OmniReason-nuScenes dataset. The complete configuration, incorporating environment description, dynamic and static objects, reasoning, and action components, achieves the strongest results, with the highest BLEU-1 score of 51.54, precision of 53.32, and recall of 55.19. Notably, the inclusion of environment description alone improves the BLEU-1 score by 7.6% compared to the baseline configuration, demonstrating the importance of contextual understanding.

ID	Environment Description	Dynamic Object	Static Scene	Reasoning	Action	VQA			Open-Loop		
						BL-1 $\uparrow$	Precision $\uparrow$	Recall $\uparrow$	L2 $\downarrow$	CR $\downarrow$	IR $\downarrow$
1		✓	✓	✓	✓	47.91	49.00	53.01	0.34	0.37	3.33
2	✓		✓	✓	✓	50.14	54.22	52.31	0.34	0.42	3.08
3	✓	✓		✓	✓	49.13	52.11	51.68	0.34	0.37	3.16
4	✓	✓	✓		✓	48.25	53.43	50.95	0.34	0.43	3.22
5	✓	✓	✓	✓		50.57	52.95	53.46	0.34	0.43	3.16
6	✓	✓	✓	✓	✓	51.54	53.32	55.19	0.34	0.40	3.18

Table 4: Ablation study on the OmniReason-nuScenes dataset. We assess OmniReason-Agent’s performance with different language components enabled by measuring VQA accuracy and open-loop planning quality, where L2 distance, collision rate, and intersection rate are evaluated as the average values over the 1, 2, and 3 s horizons.

Temporal Memory	VQA			Open-Loop		
	BL-1 $\uparrow$	Precision $\uparrow$	Recall $\uparrow$	L2 $\downarrow$	CR $\downarrow$	IR $\downarrow$
	52.24	54.84	54.86	0.38	0.44	3.65
✓	51.54	53.32	55.19	0.34	0.40	3.18

Table 5: Ablation study on sparse temporal memory module. L2 distance, collision rate, and intersection rate are evaluated as the average values over the 1, 2, and 3 s horizons.

Incorporating dynamic object descriptions is especially beneficial for safety, resulting in the lowest collision rate of 0.37. Across all configurations, L2 error remains stable at 0.34, highlighting the model’s robust trajectory prediction. Moreover, incorporating environment description and static scene analysis leads to the lowest intersection rate of 3.08, suggesting that comprehensive spatial context significantly enhances the model’s ability to anticipate and safely navigate complex intersection scenarios. These results demonstrate OmniReason-Agent’s comprehensive scene understanding capabilities, with particularly strong performance in handling dynamic scenarios and integrating environmental context. The systematic component-wise evaluation provides valuable insights into the model’s strengths across diverse driving situations and task requirements.

Effectiveness on Temporal Memory Module. Table 5 presents the results of the ablation study on the sparse temporal memory module. Introducing the temporal memory module leads to improvements in open-loop driving, as shown by decreases in L2 distance, collision rate, and intersection rate when compared to the baseline without memory. Specifically, L2 distance is reduced from 0.38 to 0.34, collision rate decreases from 0.44 to 0.40, and intersection rate drops from 3.65 to 3.18. While there is a slight reduction in VQA BL-1 and precision, the recall improves, indicating that the model becomes more sensitive to relevant temporal information. These results demonstrate that the temporal memory module enhances sequential reasoning and contributes to more accurate and safer driving performance.

Qualitative Results

Action Generation and Causal Reasoning. Fig. 4 demonstrates OmniReason’s capabilities in scene understanding,

action generation, and causal reasoning. From the results, OmniReason not only accurately perceives detailed scene information and identifies key objects that influence the behavior of the ego vehicle, but also effectively models human priors and infers appropriate motion decisions. Additional results are provided in **Appendix B**. These findings indicate that, even in extreme situations, our model consistently maintains robust performance, further highlighting its superior reasoning and decision-making abilities.

Action: Moderate Speed, Right Turn

Environment Description

Q: Describe the surrounding environment in the driving scene.

A: The vehicle is driving on a multi-lane city road with clear weather conditions, bright sunlight, and a dry road surface. The road is well-maintained with visible road markings. The vehicle is approaching an intersection, as indicated by the presence of crosswalks, traffic signs, and lane markings. The road appears to be a typical urban street with buildings, parked cars, and some greenery. The intersection has multiple lanes, including straight and turning lanes, with clear road markings and traffic signs.

Action & Trajectory

Q: Please provide the planning trajectory for the ego car without reasons.

A: Here is the planning trajectory [PT, (+2.77, -0.21), (+5.52, -0.80), (+8.35, -1.79), (+11.04, -3.14), (+13.69, -4.84), (+16.42, -6.69)].

Q: Describe the action of ego vehicle, including speed_state ("Crawling"/"Moving Slowly"/"Moderate Speed"/"Moving Fast"/"Moving Very Fast"), current_state("Vehicle Starting"/"Accelerate"/"Decelerate"/"Left Lane Changing"/"Right Lane Changing"/"Lane Keeping"). If ego vehicle is stopped, please only return "Stopped".

A: Moderate Speed, Right Turn

Causal Reasoning

Q: Give a brief rationale discussing the cause for the current action/decision.

A: The vehicle is approaching an intersection with a red light, and the decision to moderate speed and prepare for a right turn is based on the current traffic conditions. The driver needs to slow down and ensure there are no pedestrians crossing the right-front crosswalk. Checking mirrors and over the shoulder is essential for turning safely. The construction zone on the left requires attention, as it could influence traffic patterns and pedestrian behaviors. The decision to turn right is influenced by the red light ahead and the need to navigate the intersection safely.

Figure 4: Environment description, action & trajectory, and causal reasoning VQA results of OmniReason-Agent on the nuScenes dataset. The green refers to the predicted trajectory.

Conclusion

This work presents OmniReason, a comprehensive framework for advancing interpretable and reliable autonomous driving. By introducing large-scale VLA datasets with rich, temporally dense language annotations and proposing the OmniReason-Agent model that distills human-like priors and causal reasoning, we enable more context-aware and explainable driving decisions. Extensive experiments demonstrate that our approach significantly improves performance in safety, comfort, and explanation quality, bringing autonomous vehicles closer to human-level understanding.

References

- Arai, H.; Miwa, K.; Sasaki, K.; Watanabe, K.; Yamaguchi, Y.; Aoki, S.; and Yamamoto, I. 2025. CoVLA: Comprehensive Vision-Language-Action Dataset for Autonomous Driving. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision*, 1933–1943. IEEE.
- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.
- Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11621–11631.
- Chen, L.; Sinavski, O.; Hünemann, J.; Karnsund, A.; Willmott, A. J.; Birch, D.; Maund, D.; and Shotton, J. 2024a. Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving. In *2024 IEEE International Conference on Robotics and Automation*, 14093–14100. IEEE.
- Chen, L.; Wu, P.; Chitta, K.; Jaeger, B.; Geiger, A.; and Li, H. 2024b. End-to-End Autonomous Driving: Challenges and Frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Chen, S.; Jiang, B.; Gao, H.; Liao, B.; Xu, Q.; Zhang, Q.; Huang, C.; Liu, W.; and Wang, X. 2024c. VADv2: End-to-End Vectorized Autonomous Driving via Probabilistic Planning. *arXiv preprint arXiv:2402.13243*.
- Deruyttere, T.; Grujicic, D.; Blaschko, M. B.; and Moens, M.-F. 2022. Talk2Car: Predicting Physical Trajectories for Natural Language Commands. *Ieee Access*, 10: 123809–123834.
- Deruyttere, T.; Vandenhende, S.; Grujicic, D.; Van Gool, L.; and Moens, M. F. 2019. Talk2Car: Taking Control of Your Self-Driving Car. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2088–2098.
- Fang, Y.; Sun, Q.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2024. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 149: 105171.
- Fu, H.; Zhang, D.; Zhao, Z.; Cui, J.; Liang, D.; Zhang, C.; Zhang, D.; Xie, H.; Wang, B.; and Bai, X. 2025. ORION: A Holistic End-to-End Autonomous Driving Framework by Vision-Language Instructed Action Generation. *arXiv preprint arXiv:2503.19755*.
- Han, W.; Guo, D.; Xu, C.-Z.; and Shen, J. 2025. DME-Driver: Integrating Human Decision Logic and 3D Scene Perception in Autonomous Driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3347–3355.
- Hu, S.; Chen, L.; Wu, P.; Li, H.; Yan, J.; and Tao, D. 2022. ST-P3: End-to-End Vision-Based Autonomous Driving via Spatial-Temporal Feature Learning. In *European Conference on Computer Vision*, 533–549. Springer.
- Jia, X.; Yang, Z.; Li, Q.; Zhang, Z.; and Yan, J. 2024. Bench2Drive: Towards Multi-Ability Benchmarking of Closed-Loop End-To-End Autonomous Driving. *arXiv preprint arXiv:2406.03877*.
- Jiang, B.; Chen, S.; Liao, B.; Zhang, X.; Yin, W.; Zhang, Q.; Huang, C.; Liu, W.; and Wang, X. 2024. Senna: Bridging Large Vision-Language Models and End-to-End Autonomous Driving. *arXiv preprint arXiv:2410.22313*.
- Jiang, B.; Chen, S.; Xu, Q.; Liao, B.; Chen, J.; Zhou, H.; Zhang, Q.; Liu, W.; Huang, C.; and Wang, X. 2023. VAD: Vectorized Scene Representation for Efficient Autonomous Driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8340–8350.
- Kim, J.; Misu, T.; Chen, Y.-T.; Tawari, A.; and Canny, J. 2019. Grounding Human-To-Vehicle Advice for Self-Driving Vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10591–10599.
- Kim, J.; Rohrbach, A.; Darrell, T.; Canny, J.; and Akata, Z. 2018. Textual Explanations for Self-Driving Vehicles. In *Proceedings of the European Conference on Computer Vision*, 563–578.
- Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024a. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. *arXiv preprint arXiv:2407.07895*.
- Li, H.; Li, Y.; Wang, H.; Zeng, J.; Xu, H.; Cai, P.; Chen, L.; Yan, J.; Xu, F.; Xiong, L.; et al. 2023. Open-sourced Data Ecosystem in Autonomous Driving: the Present and Future. *arXiv preprint arXiv:2312.03408*.
- Li, Z.; Yu, Z.; Lan, S.; Li, J.; Kautz, J.; Lu, T.; and Alvarez, J. M. 2024b. Is Ego Status All You Need for Open-Loop End-to-End Autonomous Driving? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14864–14873.
- Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.
- Liu, M.; Yurtsever, E.; Fossaert, J.; Zhou, X.; Zimmer, W.; Cui, Y.; Zagar, B. L.; and Knoll, A. C. 2024b. A Survey on Autonomous Driving Datasets: Statistics, Annotation Quality, and a Future Outlook. *IEEE Transactions on Intelligent Vehicles*.

- Liu, Y.; Wang, T.; Zhang, X.; and Sun, J. 2022. PETR: Position Embedding Transformation for Multi-view 3D Object Detection. In *European Conference on Computer Vision*, 531–548. Springer.
- Loshchilov, I.; and Hutter, F. 2016. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv preprint arXiv:1608.03983*.
- Malla, S.; Choi, C.; Dwivedi, I.; Choi, J. H.; and Li, J. 2023. DRAMA: Joint Risk Localization and Captioning in Driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1043–1052.
- Mao, J.; Qian, Y.; Ye, J.; Zhao, H.; and Wang, Y. 2023a. GPT-Driver: Learning to Drive with GPT. *arXiv preprint arXiv:2310.01415*.
- Mao, J.; Ye, J.; Qian, Y.; Pavone, M.; and Wang, Y. 2023b. A Language Agent for Autonomous Driving. *arXiv preprint arXiv:2311.10813*.
- Marcu, A.; Chen, L.; Hünemann, J.; Karnsund, A.; Hanotte, B.; Chidananda, P.; Nair, S.; Badrinarayanan, V.; Kendall, A.; Shotton, J.; et al. 2023. LingoQA: Visual Question Answering for Autonomous Driving. *arXiv preprint arXiv:2312.14115*.
- Nie, M.; Peng, R.; Wang, C.; Cai, X.; Han, J.; Xu, H.; and Zhang, L. 2024. Reason2Drive: Towards Interpretable and Chain-Based Reasoning for Autonomous Driving. In *European Conference on Computer Vision*, 292–308. Springer.
- Pan, C.; Yaman, B.; Nesti, T.; Mallik, A.; Allievi, A. G.; Velipasalar, S.; and Ren, L. 2024. VLP: Vision Language Planning for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14760–14769.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, 8748–8763. PmLR.
- Renz, K.; Chen, L.; Marcu, A.-M.; Hünemann, J.; Hanotte, B.; Karnsund, A.; Shotton, J.; Arani, E.; and Sinavski, O. 2024. CarLLaVA: Vision language models for camera-only closed-loop driving. *arXiv preprint arXiv:2406.10165*.
- Sachdeva, E.; Agarwal, N.; Chundi, S.; Roelofs, S.; Li, J.; Kochenderfer, M.; Choi, C.; and Dariush, B. 2024. Rank2Tell: A Multimodal Driving Dataset for Joint Importance Ranking and Reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 7513–7522.
- Seff, A.; Cera, B.; Chen, D.; Ng, M.; Zhou, A.; Nayakanti, N.; Refaat, K. S.; Al-Rfou, R.; and Sapp, B. 2023. MotionLM: Multi-Agent Motion Forecasting as Language Modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8579–8590.
- Shao, H.; Hu, Y.; Wang, L.; Song, G.; Waslander, S. L.; Liu, Y.; and Li, H. 2024. LMDrive: Closed-Loop End-to-End Driving with Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15120–15130.
- Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Beißwenger, J.; Luo, P.; Geiger, A.; and Li, H. 2024. DriveLM: Driving with Graph Visual Question Answering. In *European Conference on Computer Vision*, 256–274. Springer.
- Team, Q. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.
- Tian, X.; Gu, J.; Li, B.; Liu, Y.; Wang, Y.; Zhao, Z.; Zhan, K.; Jia, P.; Lang, X.; and Zhao, H. 2024. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. *arXiv preprint arXiv:2402.12289*.
- Wang, S.; Liu, Y.; Wang, T.; Li, Y.; and Zhang, X. 2023a. Exploring Object-Centric Temporal Modeling for Efficient Multi-View 3D Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3621–3631.
- Wang, S.; Yu, Z.; Jiang, X.; Lan, S.; Shi, M.; Chang, N.; Kautz, J.; Li, Y.; and Alvarez, J. M. 2025. OmniDrive: A Holistic Vision-Language Dataset for Autonomous Driving with Counterfactual Reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22442–22452.
- Wang, W.; Xie, J.; Hu, C.; Zou, H.; Fan, J.; Tong, W.; Wen, Y.; Wu, S.; Deng, H.; Li, Z.; et al. 2023b. DriveMLM: Aligning Multi-Modal Large Language Models with Behavioral Planning States for Autonomous Driving. *arXiv preprint arXiv:2312.09245*.
- Wang, Y.; Guizilini, V. C.; Zhang, T.; Wang, Y.; Zhao, H.; and Solomon, J. 2022. DETR3D: 3D Object Detection from Multi-view Images via 3D-to-2D Queries. In *Conference on Robot Learning*, 180–191. PMLR.
- Xu, Y.; Yang, X.; Gong, L.; Lin, H.-C.; Wu, T.-Y.; Li, Y.; and Vasconcelos, N. 2020. Explainable Object-Induced Action Decision for Autonomous Vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9523–9532.
- Xu, Z.; Jain, S.; and Kankanhalli, M. 2024. Hallucination is Inevitable: An Innate Limitation of Large Language Models. *arXiv preprint arXiv:2401.11817*.
- Xu, Z.; Zhang, Y.; Xie, E.; Zhao, Z.; Guo, Y.; Wong, K.-Y. K.; Li, Z.; and Zhao, H. 2024. DriveGPT4: Interpretable End-to-End Autonomous Driving Via Large Language Model. *IEEE Robotics and Automation Letters*.
- Yang, J.; Gao, S.; Qiu, Y.; Chen, L.; Li, T.; Dai, B.; Chitta, K.; Wu, P.; Zeng, J.; Luo, P.; et al. 2024. Generalized Predictive Model for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14662–14672.
- Yuan, J.; Sun, S.; Omeiza, D.; Zhao, B.; Newman, P.; Kunze, L.; and Gadd, M. 2024. RAG-Driver: Generalisable Driving Explanations with Retrieval-Augmented In-Context Learning in Multi-Modal Large Language Model. *arXiv preprint arXiv:2402.10828*.
- Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid Loss for Language Image Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11975–11986.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; et al. 2025. Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models . *Computational Linguistics*, 1–45.

Zhou, X.; Han, X.; Yang, F.; Ma, Y.; and Knoll, A. C. 2025. OpenDriveVLA: Towards End-to-end Autonomous Driving with Large Vision Language Action Model. *arXiv preprint arXiv:2503.23463*.

Zhou, X.; and Knoll, A. C. 2024. GPT-4V as Traffic Assistant: An In-depth Look at Vision Language Model on Complex Traffic Events. *arXiv preprint arXiv:2402.02205*.

Zhou, X.; Liu, M.; Yurtsever, E.; Zagar, B. L.; Zimmer, W.; Cao, H.; and Knoll, A. C. 2024. Vision Language Models in Autonomous Driving: A Survey and Outlook. *IEEE Transactions on Intelligent Vehicles*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*.

A OmniReason-Data

A.1 Scene-Aware Spatial Captioning

Manual caption annotation for VLA datasets in autonomous driving is prohibitively time-consuming, costly, and not feasible for large-scale deployment. Meanwhile, relying solely on standard large model-based captioning can lead to hallucinated and weakly grounded descriptions, which undermines dataset reliability. To address these challenges, as illustrated in Fig. 1, we propose an automated spatial captioning pipeline that begins with a rule-based annotation stage. Here, structured templates leverage rich scene cues, such as map context, vehicle dynamics, trajectory profiles, surrounding agents, and traffic signals, to generate detailed, context-aware captions for every video frame. This process ensures scalability, consistency, and robust grounding, while substantially reducing hallucination compared to naïve large-model captioning, thus laying a reliable foundation for large-scale autonomous driving datasets.

A.2 Human-Prior Knowledge Guided

To further enhance the interpretability and human-likeness of the generated captions, we introduce a human-prior knowledge guided stage. This component incorporates explicit prompts centered around driving rules, visual reasoning principles, and scenario-specific heuristics, as shown at the top of Fig. 1. By embedding these knowledge-driven cues, the captioning process is steered toward producing language and rationales that reflect the thought process and intuition of an experienced human driver. This augmentation allows the dataset to capture high-level explanations, intention recognition, and structured abstraction that go beyond the descriptive capability of rule-based templates alone, resulting in more instructive and contextually aware supervision for VLA research.

A.3 MLLM-based Temporal Reasoning

Building on the robust grounding provided by scene-aware spatial annotations and human-prior knowledge cues, we enhance temporal coherence and causal understanding through MLLM-based captioning. As highlighted in Fig. 1, both the scene-aware spatial ground truth and Human-Prior Knowledge are formatted as structured prompts and fed directly into the Qwen2.5-VL-72B model alongside multi-view, multi-frame video segments. Unlike single-frame captioning solutions such as OpenDV-2K or BLIP-2, our method operates on 5-frame video windows, capturing the temporal evolution of the environment, agent interactions, and sequential decision-making processes that are critical for autonomous driving. For each 30-second driving scene, the video is partitioned into overlapping temporal windows, comprehensively sampling both spatial and temporal dynamics, and providing rich contextual information for the MLLM to reason over. This design allows the model to generate captions that not only describe the current state of the environment but also integrate causal reasoning, future intent, and contextual understanding derived from human-like prompts and scene-aware cues. Leveraging large-scale distributed inference with 128 NVIDIA H20 GPUs, our pipeline effi-

ciently produces over 22.7w temporally grounded, causally enriched, and linguistically diverse captions within a single day, dramatically scaling up high-quality annotation resources while faithfully reflecting the complexity and reasoning demands of real-world driving scenarios.

A.4 Data Analysis

Environment Analysis. Due to the large data volume and high diversity of Bench2Drive, we primarily showcase analysis on the Bench2Drive dataset. Fig. 5 and Fig. 6 illustrate the structural diversity and complexity of road layouts within Bench2Drive. As shown in Fig. 5, the dataset encompasses a wide range of lane configurations, with most scenes featuring one or two same-direction lanes and varying numbers of opposite-direction lanes, leading to total lane counts that are predominantly two or three, but also covering higher numbers. This reflects a substantial spectrum of real-world driving environments. Fig. 6 further investigates the occurrence of cross lanes, revealing that approximately half of the samples contain intersecting lanes, thereby capturing both regular and complex intersection scenarios. The breakdown of cross lane types, including left-to-right and right-to-left vertical crossings, highlights the presence of multifaceted intersection designs. Collectively, these characteristics demonstrate that Bench2Drive offers rich and challenging road structures, supporting comprehensive evaluation of autonomous driving models across a diverse range of traffic scenarios. Fig. 7 presents the joint distribution of the number of surrounding vehicles and pedestrians per frame alongside their minimum distance to the ego vehicle, offering a comprehensive view of both scene density and interaction complexity. The left subplot demonstrates that frames with a larger number of vehicles are associated with reduced minimum distances, reflecting a diverse spectrum of dense and potentially challenging traffic scenarios where close-proximity interactions are common. In contrast, the right subplot shows that while most frames feature zero or one pedestrian, there is still a notable range of cases with multiple pedestrians at varying distances, capturing complex urban scenes where pedestrian interactions can occur. Overall, these results illustrate that the dataset encompasses a wide variety of real-world driving situations, from highly dynamic multi-vehicle environments to scenes with sparse or clustered pedestrian activity, thus supporting robust evaluation of autonomous driving systems under both common and challenging conditions. relatively rare, with most occurring at greater distances from the ego vehicle.

Action Analysis. Fig. 8 and Fig. 9 illustrate the distribution of ego vehicle future actions along two key dimensions: the high-level longitudinal action category (such as accelerating, decelerating, or vehicle starting) and the fine-grained maneuver type (such as lane keeping, lane changing, and turning), conditioned on the current driving state.

Across both the nuScenes and Bench2Drive datasets, similar trends are observed. When the ego vehicle is in the Crawling state, Vehicle Starting is the most common future action, reflecting typical behavior as vehicles transition from a stop. As the current state shifts to Moderate Speed, future

actions become more evenly distributed, with Accelerate and Decelerate gaining prominence, indicative of dynamic adjustments in flowing traffic. In the Moving Fast state, the repertoire of future actions broadens, yet the incidence of Vehicle Starting diminishes significantly in both datasets.

The right panels of each figure provide further granularity, showing distributions of specific maneuvers. Here, Go Straight overwhelmingly dominates at moderate and high speeds, while more complex behaviors such as lane changes and turns are less frequent. Notably, the Bench2Drive dataset exhibits a higher proportion of lane change and turning maneuvers, particularly in low-speed states, suggesting greater environmental diversity and more elaborate behavior annotations compared to nuScenes. Overall, these results demonstrate both consistent trends and meaningful differences across datasets, highlighting the diversity of scene contexts and the complexity of decision-making captured within each.

Causal Reasoning Analysis Fig. 10 presents word clouds of causal reasoning annotations for the nuScenes (left) and Bench2Drive (right) datasets, highlighting the core concepts and decision factors that guide autonomous driving behavior in each setting. Prominent terms such as "vehicle," "presence," and "decision" appear in both datasets, reflecting a shared emphasis on situational awareness and decision-making. Additional keywords like "pedestrian," "traffic light," "moderate speed," and "safe distance" reveal considerations of dynamic agents, traffic regulations, and safety constraints. The distribution and diversity of these terms illustrate the contextual richness and the range of challenges addressed by causal reasoning in complex real-world scenarios across both datasets.

Case Study. Fig. 11 and Fig. 12 from the OmniReason-Bench2Drive dataset showcase the annotation model’s scene understanding capability and ability to reason effectively in complex and challenging driving scenarios. Fig. 11 depicts a scenario where the ego vehicle navigates a foggy two-lane city road as a pedestrian crosses its path, requiring the vehicle to proceed slowly and remain ready to stop. Additional dynamic agents, such as a following vehicle and an oncoming car, further complicate the scene. Fig. 12 presents a rural road during twilight with wet conditions, reduced visibility, and multiple hazards, including oncoming traffic and cyclists. In both cases, the lack of traffic signals and the presence of dynamic obstacles highlight the importance of cautious, context-aware planning. These examples demonstrate the dataset’s capacity to capture realistic challenges and support the development of interpretable, safety-focused autonomous driving models.

B Additional Results

B.2 Ablation Study

Generalizability on OmniDrive Dataset. Table 6 presents a comprehensive evaluation of OmniReason-Agent against state-of-the-art multimodal models on the OmniDrive dataset. Our model establishes new benchmarks across all evaluation metrics, achieving a CIDEr score of

86.67, representing a 3.0% improvement over the previous best model LLaVa-next 72B. The BLEU-1 score of 36.40 demonstrates an even more substantial 7.1% gain, while the BLEU-4 score shows a remarkable 42.4% improvement. The performance advantage is particularly notable when comparing models of similar scale. Our model outperforms Qwen2.5VL 3B by 18.1% in CIDEr and 23.0% in BLEU-1, despite identical model sizes. More significantly, our solution surpasses much larger 72B parameter models, exceeding Qwen2.5VL 72B by 7.7% in CIDEr and LLaVa-next 72B by 3.0% in the same metric. This comprehensive performance advantage demonstrates that OmniReason-Agent’s architectural innovations deliver superior results regardless of model scale, establishing new state-of-the-art performance for autonomous driving applications.

Model	CI-r	BL-1	BL-4	ME-R	RO-L
Qwen2.5VL 72B	80.45	32.96	4.46	40.41	23.56
Qwen2.5VL 32B	76.13	25.86	2.82	29.36	20.77
Qwen2.5VL 7B	76.93	32.46	3.95	44.46	23.11
Qwen2.5VL 3B	73.41	29.59	3.50	46.70	23.48
LLava-next 72B	84.14	33.98	5.28	42.44	23.87
OmniReason-Agent	86.67	36.40	7.52	47.58	27.87

Table 6: Performance comparison on the OmniDrive dataset (Wang et al. 2025). OmniReason-Agent outperforms prior works in all metrics.

B.1 Qualitative Results

Fig. 13 and Fig. 14 showcase the VQA results of OmniReason-Agent on the nuScenes dataset for two distinct urban intersection scenarios. In Fig. 13, the ego vehicle identifies the presence of pedestrians crossing from both directions and makes an interpretable, safety-oriented decision to stop, demonstrating strong scene understanding and adherence to traffic rules. Fig. 14 presents a left turn scenario where the agent recognizes a green traffic light and clear road markings, and cautiously executes a slow left turn while monitoring a vehicle ahead. Together, these examples highlight the model’s ability to integrate static and dynamic scene information, perform causal reasoning, and generate reliable, context-aware actions in complex real-world traffic situations.

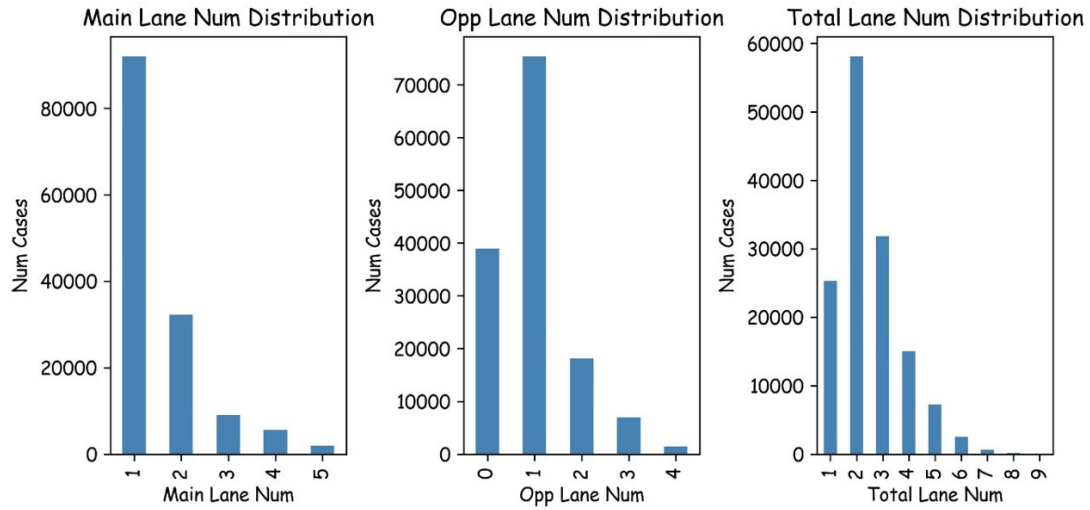


Figure 5: Lane number distributions in the Bench2Drive dataset. The left plot shows the distribution of main (ego-moving direction) lane numbers, with the majority of cases having one or two same-direction lanes. The middle plot displays the distribution of opposite-direction lanes, where most scenarios involve a single or no opposite lane. The right plot shows the total number of lanes, indicating that most roads in the dataset have two to three lanes in total.

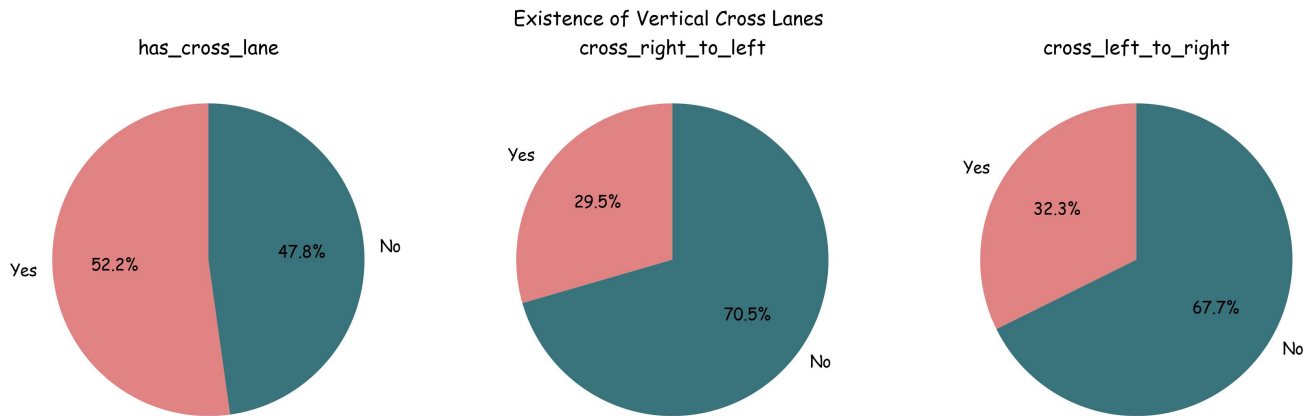


Figure 6: Distribution of cross lanes in the Bench2Drive dataset. The left pie chart shows the proportion of samples with cross lanes (has_cross_lane). The middle and right pie charts illustrate the existence of vertical cross lanes from right to left (cross_right_to_left) and from left to right (cross_left_to_right).

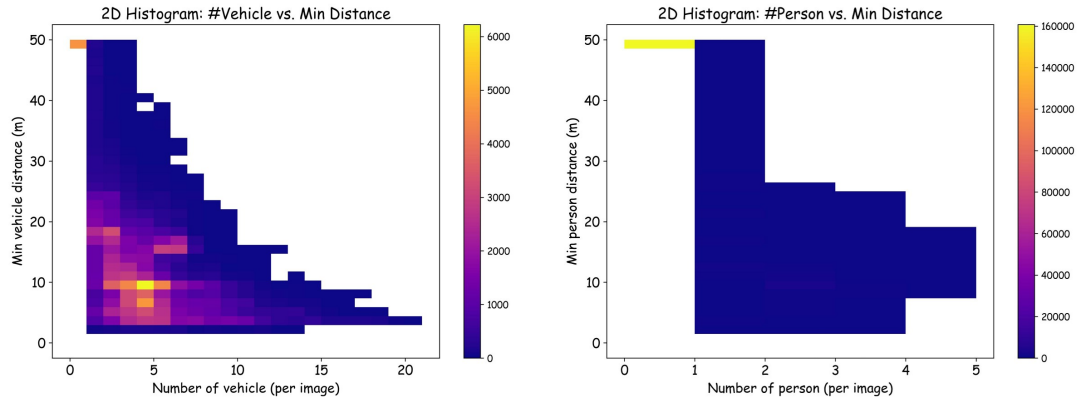


Figure 7: 2D histograms visualizing the relationship between the number of objects and their minimum distance to the ego vehicle in each image. The left plot shows the distribution for vehicles, indicating that as the number of vehicles in the scene increases, the minimum distance to the nearest vehicle generally decreases. The right plot shows the distribution for pedestrians, where most images contain zero or one person, and the minimum distance to the nearest person is typically large. The color bars represent the number of cases in each bin.

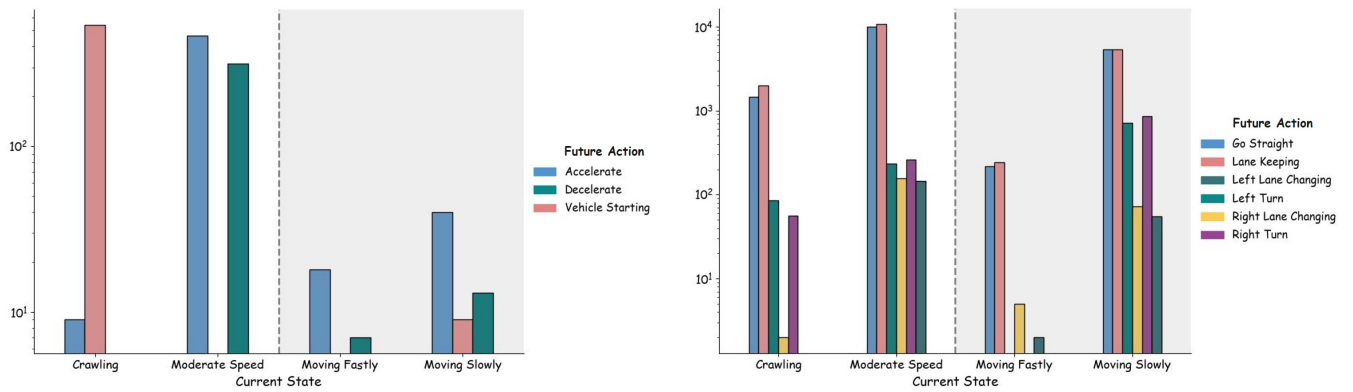


Figure 8: Ego vehicle's future action at different current states on the nuScenes dataset.

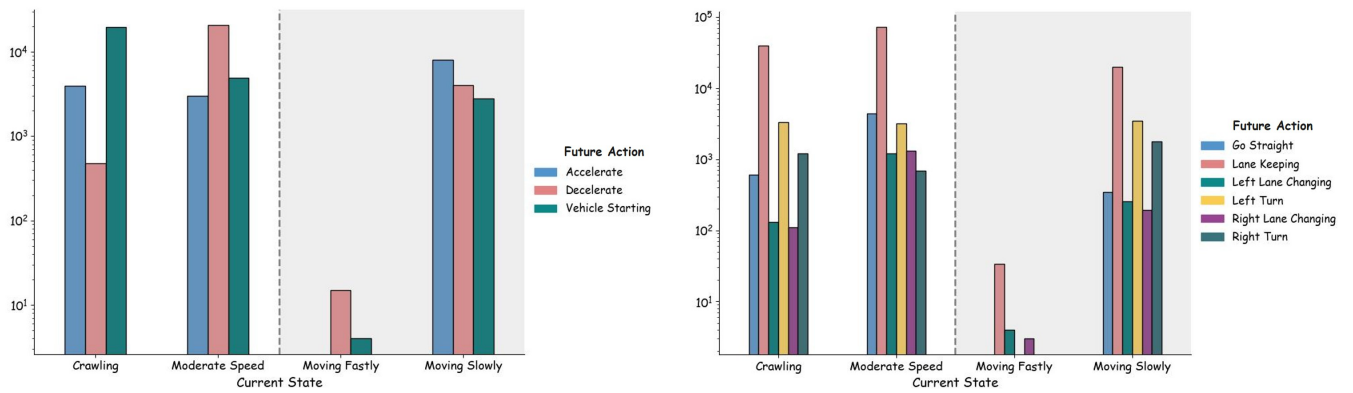


Figure 9: Ego vehicle's future action at different current states on the Bench2Drive dataset.



Environment Description

"The vehicle is on a two-lane city road with light fog conditions, reducing visibility slightly. The road appears to be well-maintained with clear lane markings. The lighting suggests daytime with no direct sunlight due to the fog. There are trees and rocks lining the sides of the road, and a food stall is visible on the right side. The road does not appear to be approaching an intersection, as there are no visible intersection markings or traffic lights. The road markings indicate a straight path ahead."

Dynamic Object Analysis

"A pedestrian is crossing the road ahead at coordinates (+2.7, +0.3) with a velocity of (+0.4, +2.0), indicating they are moving forward and slightly towards the left. This pedestrian poses a potential risk as they are within the vehicle's path. A car is positioned behind the vehicle at coordinates (-16.5, -0.5) with a velocity of (+8.9, +0.3), suggesting it is following at a moderate speed. Another car is visible in the opposite lane at coordinates (+11.8, +4.5) with a velocity of (+16.4, +0.5), traveling in the opposite direction and not posing an immediate threat."

Static Scene Analysis

"The road has two lanes: one for the vehicle's direction and one for oncoming traffic. There are no visible traffic lights, stop lines, or crosswalks in the immediate vicinity. The lane markings are clear, and the road appears to be relatively straight without any sharp turns. The presence of a food stall on the right side of the road might attract more pedestrians, so extra caution is needed in that area."

Causal Reasoning

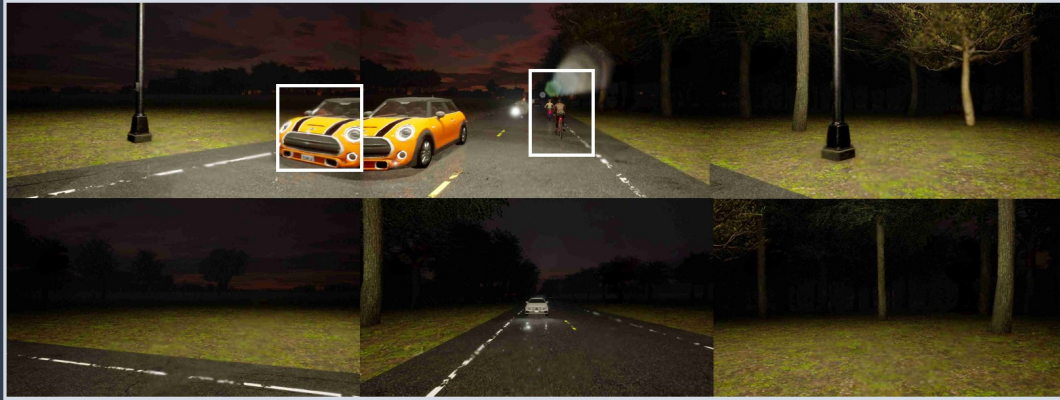
"Given the pedestrian crossing the road ahead, the vehicle is crawling and preparing to start while maintaining lane keeping. The pedestrian's proximity requires the vehicle to remain cautious and ready to stop if necessary. The car behind is at a safe distance, and the oncoming traffic is not an immediate concern. The light fog reduces visibility, so the vehicle should proceed slowly and cautiously to ensure safety."

Action

Expert decision: Crawling, Vehicle Starting, Lane Keeping.

Trajectory

[PT, (-1.33, -0.07), (-0.58, +0.51), (-1.27, +0.82), (-1.06, -1.27), (-1.33, -0.26), (-1.50, -0.66)]



Environment Description

"The vehicle is driving on a two-lane rural road during twilight hours, with heavy rain and light fog reducing visibility. The road appears wet due to the rain, and the sky is partly cloudy. The road markings are visible but slightly obscured by the weather conditions. There are no intersections or traffic signals in sight, and the road is flanked by trees and grassy areas. The lighting is dim, primarily coming from the vehicle's headlights and ambient light reflecting off the wet road surface."

Dynamic Object Analysis

"A car is visible ahead on the left side of the road, traveling in the opposite direction. Its headlights are on, indicating it is moving towards the vehicle. Additionally, there are two cyclists riding ahead on the right side of the road, moving in the same direction as the vehicle. Their presence requires careful maneuvering to maintain a safe distance. No other dynamic objects are immediately threatening or close enough to affect the driving behavior."

Static Scene Analysis

"The road has two lanes separated by a dashed yellow line, indicating a two-way traffic flow. There are no visible traffic signs, traffic lights, or stop lines within the immediate vicinity. The road edges are marked by solid white lines. The surrounding area consists of trees and grass, with no significant static obstacles like cones or barriers affecting the driving path."

Causal Reasoning

"Given the weather conditions and reduced visibility, the vehicle is moving slowly and maintaining lane keeping. The presence of oncoming traffic and cyclists necessitates cautious driving to avoid collisions. The decision to move slowly ensures better control and reaction time in case of sudden changes in the environment or actions by other road users. The absence of intersections or traffic signals allows for a straightforward approach, focusing on maintaining a safe distance from the cyclists and the oncoming car."

Action

Expert decision: Moving Slowly, Vehicle Starting, Lane Keeping.

Trajectory

[PT, (-2.98, +0.88), (-2.38, +0.27), (-1.66, -0.03), (-1.19, +2.06), (+0.72, +1.05), (+3.20, +1.46)]

Figure 12: A hazard at side lane two ways example from the OmniReason-Bench2Drive dataset is illustrated. The white box represents the dynamic objects around the ego vehicle.

Action: Stopped

Environment Description

Q: Describe the surrounding environment in the driving scene.

A: The vehicle is driving on a multi-lane city road with clear weather conditions, bright sunlight, and a dry road surface. The road is well-maintained with visible road markings and traffic signs. The vehicle is approaching an intersection, as indicated by the presence of crosswalks. The road is bordered by sidewalks, trees, and urban buildings, suggesting a typical urban environment. The road markings include lanes for straight and turning movements, and the intersection is marked with **pedestrian crossings**.

Dynamic Object Analysis

Q: Analyze dynamic object that may affect the driving.

A: There are several dynamic objects in the scene that may affect driving. **A pedestrian is moving in the front left, heading from left to right**, and another pedestrian is moving in the rear right, heading from right to left. A moving vehicle is present in the rear left, traveling in the same direction as the vehicle. These objects require attention to ensure safe navigation, especially when turning or approaching the intersection. The pedestrian's proximity to the vehicle necessitates extra caution to avoid any potential conflicts.

Static Scene Analysis

Q: Analyze static scenarios that may affect driving.

A: The static elements in the scene include multiple lanes, crosswalks. The vehicle is currently in a straight lane. There are crosswalks at the intersection. The road markings are clear, with lane dividers and crosswalk lines visible.

Causal Reasoning

Q: Give a brief rationale discussing the cause for the current action/decision.

A: The vehicle is currently stopped because there are pedestrians crossing the intersection ahead. The presence of crosswalks and pedestrians moving both from left to right and right to left requires the vehicle to yield and ensure their safety before proceeding. Stopping at this moment prevents potential conflicts and aligns with traffic regulations at pedestrian crossings.

Action & Trajectory

Q: Please provide the planning trajectory for the ego car without reasons.

A: Here is the planning trajectory [PT, (0.0, 0.0), (0.0, 0.0), (0.0, 0.0), (0.0, 0.0), (0.0, 0.0), (0.0, 0.0)].

Q: Describe the action of ego vehicle, including speed_state ("Crawling"/"Moving Slowly"/"Moderate Speed"/"Moving Fastly"), current_state("Vehicle Starting"/"Accelerate"/"Decelerate"), lane_change("Left Lane Changing"/"Right Lane Changing"/"Lane Keeping"). If ego vehicle is stopped, please only return "Stopped".

A: Stopped

Figure 13: A stop VQA results of OmniReason-Agent on the nuScenes dataset.

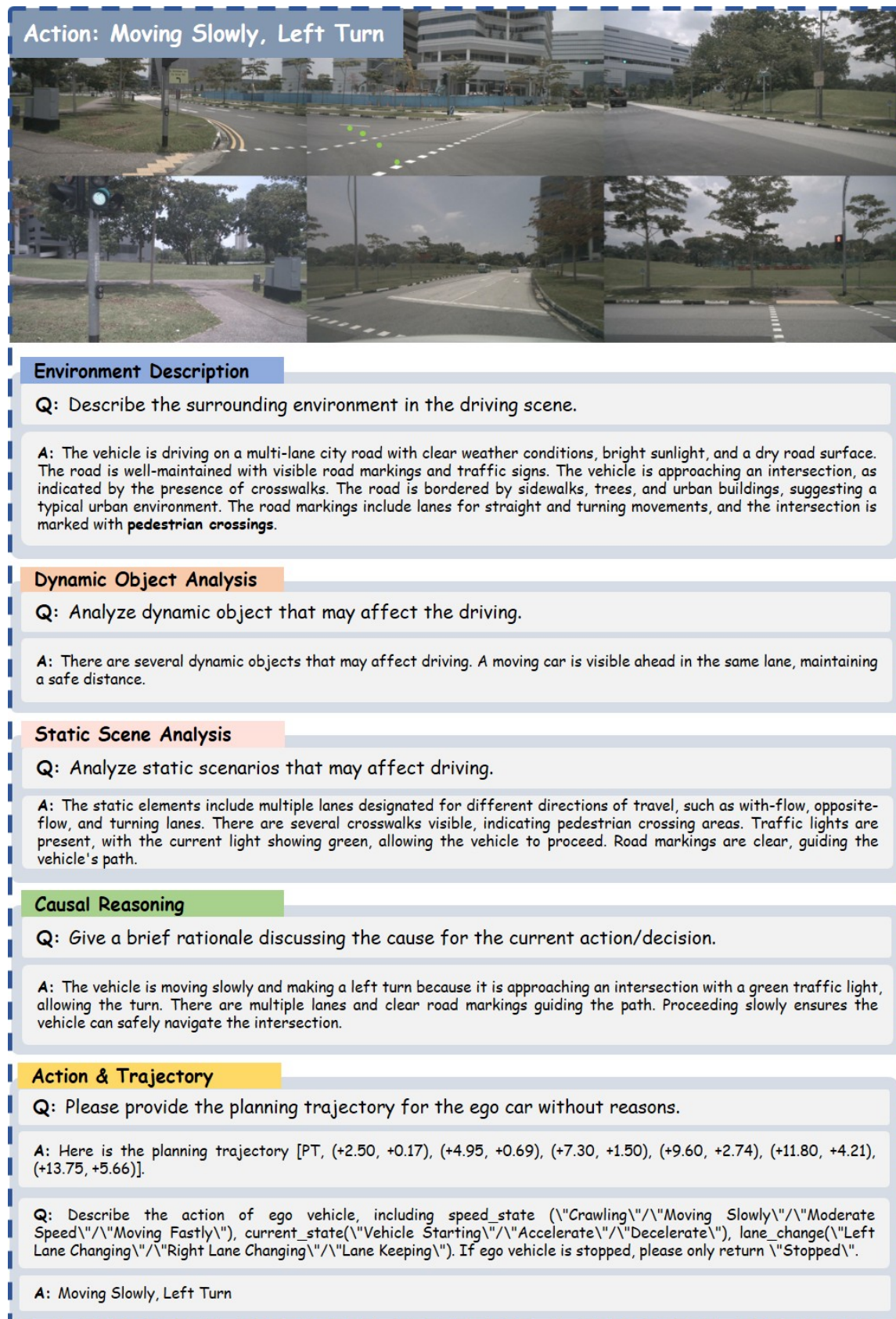


Figure 14: A left turn VQA results of OmniReason-Agent on the nuScenes dataset. The green refers to the predicted trajectory.