

CogDriver: Integrating Cognitive Inertia for Temporally Coherent Planning in Autonomous Driving

Pei Liu^{1,2*} Qingtian Ning^{3*} Xinyan Lu³ Haipeng Liu³ Weiliang Ma^{3†} Dangen She³
Xianpeng Lang³ Jun Ma^{1,2‡}

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

²The Hong Kong University of Science and Technology ³Li Auto Inc.

pliu061@connect.hkust-gz.edu.cn, jun.ma@ust.hk

Abstract

The pursuit of autonomous agents capable of temporally coherent planning is hindered by a fundamental flaw in current vision-language models (VLMs): they lack cognitive inertia. Operating on isolated snapshots, these models cannot form a continuous understanding of the environment, leading to erratic decision jitter and a failure to execute complex, multi-step maneuvers. To remedy this, we introduce CogDriver, a framework designed to build a stable internal representation by instilling this crucial cognitive property. Our work makes two key contributions: (1) We present CogDriver-Data, a large-scale vision-language-action dataset whose narrative annotations provide the supervisory signal for learning temporal dynamics and persistent intent. (2) We develop the CogDriver-Agent, an architecture featuring a sparse temporal memory to maintain a stable internal state. This is enabled by a spatiotemporal knowledge distillation approach that explicitly teaches decision coherence. Comprehensive experiments validate our paradigm: CogDriver-Agent achieves a 22% increase in the closed-loop Driving Score on Bench2Drive and a 21% reduction in mean L2 error on nuScenes, establishing a new state-of-the-art. These significant gains in both long-term decision-making and imitation accuracy provide strong evidence that our agent successfully maintains a temporally coherent internal state, bridging the gap toward more reliable autonomous driving. Project link: [CogDriver](#).

1. Introduction

The grand challenge of autonomous driving is not merely to create a system that perceives, but one that understands and anticipates the world through stable, temporally coherent reasoning [8, 45, 49]. While much of the autonomous driving literature focuses on "world models" that explicitly

predict future states in pixel or latent space, a foundational prerequisite for such anticipation is the ability to maintain a consistent internal representation over time. However, a fundamental and often overlooked flaw plagues current end-to-end systems: they operate as cognitively stateless agents, trapped in a perpetual present [4, 51]. Lacking a continuous thread of thought, they are incapable of building a stable internal state; instead, they process each moment as an isolated problem, akin to an amnesiac driver re-evaluating the world from scratch every fraction of a second.

The advent of large vision-language models (VLMs) presents a tantalizing opportunity to imbue driving systems with advanced reasoning capabilities [22, 50]. Yet, naively applying them often exacerbates the problem. Current VLMs suffer from a lack of cognitive inertia, the natural persistence of intentions that underpins stable human cognition [31, 48]. This deficit stems from a more fundamental failure: the inability to maintain temporal coherence. Without this cognitive anchor, an agent's internal representation is fragmented and ephemeral. This manifests externally as debilitating decision jitter and a failure to execute complex, multi-step maneuvers. Consider, for instance, the complex decision of overtaking a slow-moving truck on a multi-lane highway. An agent without cognitive inertia might initially decide to overtake from the left. However, upon detecting a fast-approaching vehicle in the left lane for a brief moment, it might abruptly cancel the maneuver and swerve back, only to then reconsider overtaking from the right as that lane momentarily clears. This erratic oscillation between competing strategies, "overtake left," "abort," "overtake right", makes the vehicle dangerously unpredictable. In contrast, a human driver forms a stable, multi-step plan, such as: "The left lane is too risky; I will wait for the car on the right to pass, then execute a clean overtake from the right." This commitment to a chosen strategy, guided by a coherent understanding of the environment, is not a luxury; it is a fundamental prerequisite

*Equal contribution. †Project leader. ‡Corresponding author.

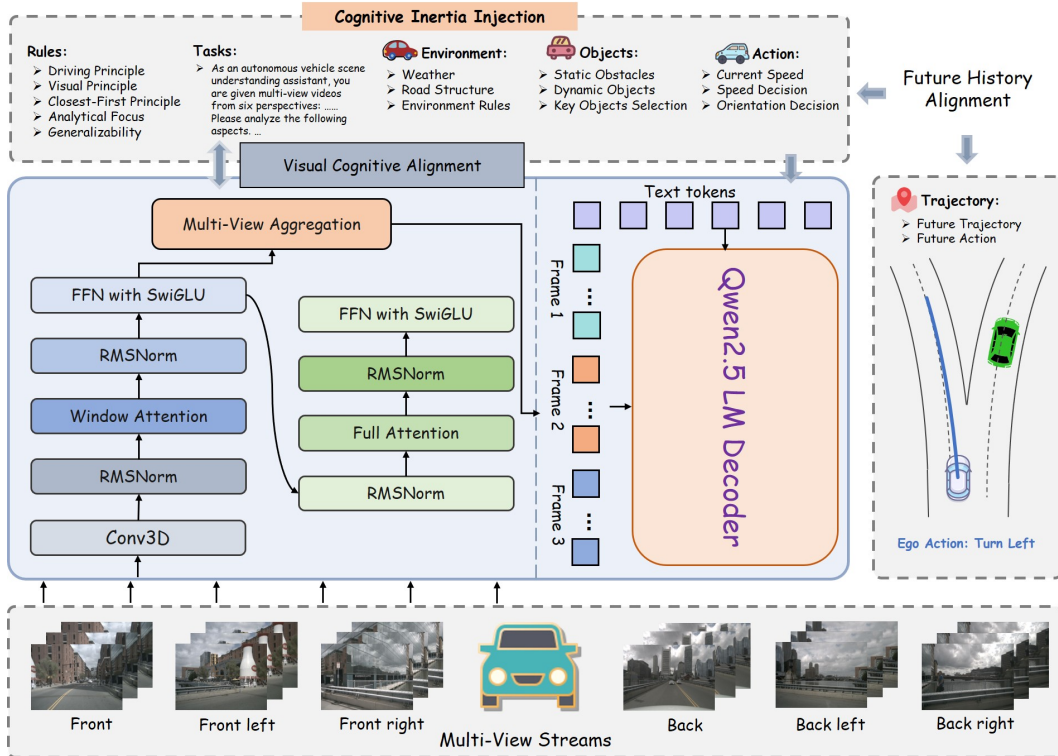


Figure 1. Annotation pipeline. To generate temporally coherent data, we propose a novel Multi-View Spatiotemporal MLLM capable of processing concurrent video streams. Its reasoning is guided by our cognitive inertia Injection framework, which provides structured rules and tasks. The generated narrative is then rigorously verified against the ground-truth vehicle trajectory via Future History Alignment, ensuring the final annotations are both causally sound and coherent.

for safe and trustworthy interaction in a dynamic world.

This leads to a more fundamental research question: **How can we build VLA agents that develop a coherent internal representation, enabling them to act with the stability and foresight of a human?**

To answer this, we introduce CogDriver, a framework designed not merely to process temporal data, but to lay the groundwork for temporally coherent planning by instilling cognitive inertia. Our work pioneers this direction through two key contributions. First, we present CogDriver-Data, two large-scale vision-language-action (VLA) datasets. Unlike existing datasets with disconnected rationales, CogDriver-Data’s narrative annotations capture the story of persistent driving intent. This narrative structure provides the crucial supervisory signal for learning the temporal dynamics essential for stable planning.

Second, building on this foundation, we develop the CogDriver-Agent. Its architecture is designed to maintain a stable internal state, a prerequisite for consistent decision-making. It features a sparse temporal memory module that acts as a mechanism for intentional persistence, effectively forming a coherent temporal context over time. This is enabled by our spatiotemporal knowledge distillation ap-

proach, which explicitly teaches the model to maintain decision coherence by learning from the narrative structure of CogDriver-Data.

Comprehensive experiments show that CogDriver-Agent not only achieves state-of-the-art performance but, more critically, exhibits significantly reduced decision jitter and successfully executes long-term plans. These emergent capabilities serve as strong evidence that our agent is developing a more stable internal representation. This work represents a paradigm shift from building reactive predictors to engineering cognitively coherent agents, marking a crucial step towards trustworthy autonomous systems powered by robust temporal reasoning.

Our contributions are summarized as follows:

- We present CogDriver-Data, two large-scale VLA datasets with novel narrative annotations designed to capture persistent intent and temporal coherence, providing the first benchmark for training and evaluating cognitively coherent agents.
- We propose CogDriver-Agent, an architecture designed to build a coherent internal representation by instilling cognitive inertia. It features a sparse temporal memory to maintain a stable internal state, trained via a knowledge

distillation method that promotes decision coherence.

- We demonstrate through extensive experiments that our approach not only sets a new state-of-the-art in standard benchmarks but also quantifiably reduces decision jitter and enables the execution of complex, long-term plans, validating the effectiveness of our paradigm.

2. Related Work

2.1. Language-enhanced Dataset for Autonomous Driving

The evolution of driving datasets has recently shifted towards incorporating language to explain actions [19, 22]. However, existing works fundamentally fail to capture cognitive inertia, treating driving as a series of disconnected moments. Pioneering datasets like BDD-X [17] and DRAMA [25] offer only atomic, snapshot-based rationales. While subsequent works like DriveLM [36] and CoVLA [1] introduced continuous trajectories, they still lack the corresponding "continuous why", the evolving causal narrative that connects decisions over time. Thus, no existing dataset provides the essential ingredients to learn cognitive inertia: the joint modeling of continuous action, holistic multi-view perception, and a persistent, evolving thought process. A thorough comparison between existing prompt-based driving datasets and ours is summarized in Table 1.

2.2. MLLMs Grounded Autonomous Driving

Recent multimodal large language models (MLLMs) for driving, despite diverse architectures like prompt-based planners [3, 26], end-to-end frameworks [44], and hybrid designs [32, 35, 41], largely operate as stimulus-response mechanisms. They fundamentally fail to model the causal relationships within dynamic scenes, a critical capability for explainable decision-making. Our work addresses this gap by proposing a framework that explicitly reasons over spatiotemporal causal dependencies. This approach unlocks zero-shot E2E planning, advancing beyond the conventional stimulus-response mappings of prior studies [9, 27, 28, 30, 34, 46] by grounding decisions in causal-temporal principles.

3. CogDriver-Data

We introduce CogDriver-Data, a new suite of large-scale VLA datasets, CogDriver-nuScenes, and CogDriver-Bench2Drive, designed to address a critical gap in existing autonomous driving data: the lack of narrative coherence. While prior datasets provide per-frame rationales, CogDriver-Data features annotations that tell a continuous story of driving decisions, capturing persistent intent, causal reasoning, and the resulting actions. As illustrated in Fig. 1, these narrative-rich annotations are generated by a novel

pipeline designed to instill spatiotemporal and logical consistency, providing a robust foundation for training cognitively coherent agents.

3.1. Data Annotation Pipeline

Our annotation pipeline is built upon two core technical innovations designed to generate data with unprecedented spatiotemporal richness and logical consistency.

First, we propose a novel Multi-View Spatiotemporal MLLM (MVST-MLLM) architecture as the annotator. Unlike prior works that process static images or single video streams, our model’s vision encoder is the first of its kind to process multi-view streams concurrently. It employs a hierarchy of Conv3D and window attention layers to extract and fuse features across both space (all camera views) and time. This holistic perception is critical, enabling the model to reason about complex dynamic events that are only comprehensible by correlating information from multiple viewpoints simultaneously, e.g., a car merging from the right while a pedestrian appears on the left.

Second, we introduce cognitive Inertia injection, a specific mechanism that enforces temporal coherence. This is not a vague concept but a direct alignment of static rules with dynamic visual evidence. The MLLM is conditioned on a structured prompt containing high-level Rules and Tasks. The core innovation lies in training the model to use these static principles to generate a single, continuous narrative that explains the entire temporal sequence of visual inputs. This alignment compels the model to produce causally-linked explanations rather than disconnected, per-frame descriptions. This process is further grounded by Future History Alignment, which verifies the generated narrative against the actual vehicle trajectory, ensuring physical plausibility.

The resulting CogDriver-Data is therefore a direct product of this architecture’s unique ability to see in 360 degrees over time and its training to bind static rules to dynamic events, creating a robust foundation for our agent. Further details are provided in the Appendix.

4. CogDriver-Agent

We present CogDriver-Agent, an end-to-end framework engineered to instill cognitive inertia by seamlessly integrating spatiotemporal perception with language-conditioned planning. As depicted in Fig. 2, the agent’s architecture is not a simple pipeline but a synergistic system. Its perceptual foundation is a hierarchical vision backbone that distills raw multi-view inputs into a compact set of spatiotemporal tokens. Central to this process is our Temporal Coherence Module (TCM), which maintains a dynamic, memory-efficient representation of the world state, enabling robust long-range reasoning. These distilled world tokens are then projected into the embedding space of a frozen VLM. Here,

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.

Dataset	Vision Data		Language Data		Action Data	Reason	Cogni	Weat	Road	Trfc.	Trfc.	Sur.	Obj.
	M.V.	Temporal	VQA	Auto.	Type	-ing	-tive	-her	Type	Light	Sign	Obj.	Pos.
Talk2Car [5]	✗	✗	✗	✗	Traj.	✗	✗	✗	✗	✗	✗	✓	✗
T2C-Traj [6]	✗	✗	✗	✗	Traj., Command	✗	✗	✗	✗	✗	✗	✓	✗
DriveLM-nuScenes [36]	✗	✗	✓	✗	Traj.	✓	✗	✗	✗	✗	✗	✓	✗
DriveLM-CARLA [36]	✓	✗	✓	✓	Traj.	✓	✗	✗	✗	✗	✗	✓	✗
DRAMA [25]	✗	✓	✓	✓	Command	✓	✗	✗	✓	✓	✓	✓	✗
Rank2Tell [33]	✗	✓	✗	✓	Command	✓	✗	✗	✗	✓	✗	✓	✓
Reason2Drive [29]	✗	✗	✓	✓	Command	✓	✗	✗	✗	✗	✗	✓	✗
CoVLA [1]	✗	✗	✗	✓	Traj.	✗	✗	✗	✓	✓	✗	✓	✗
Omnidrive [40]	✓	✗	✓	✓	Traj.	✓	✓	✓	✗	✓	✓	✓	✓
CogDriver-nuScenes (Ours)	✓	✓	✓	✓	Traj., Command	✓	✓	✓	✓	✓	✓	✓	✓
CogDriver-Bench2Drive (Ours)	✓	✓	✓	✓	Traj., Command	✓	✓	✓	✓	✓	✓	✓	✓

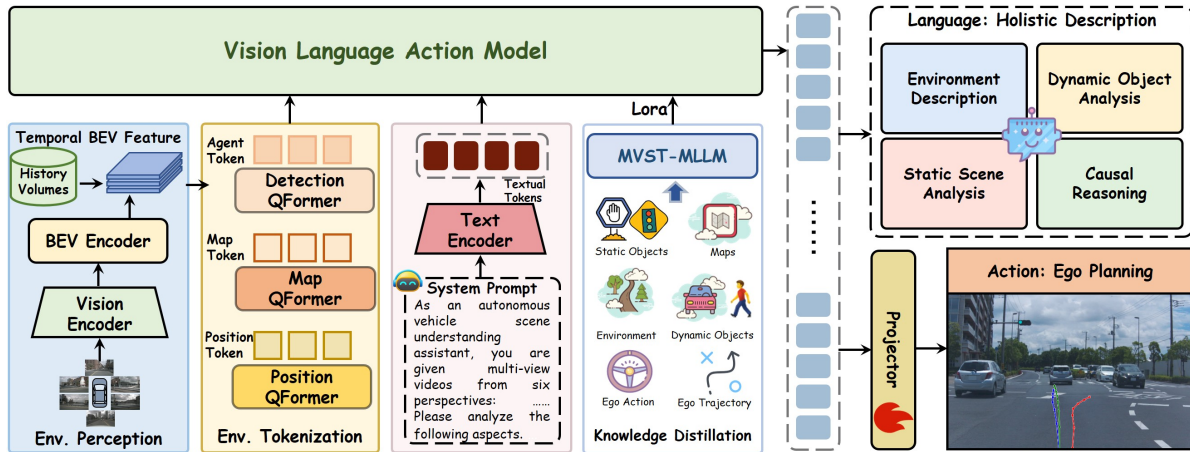


Figure 2. An overview of our CogDriver-Agent. It moves beyond reactive decision-making by harnessing a pre-trained language model to maintain cognitive inertia. It achieves this by building a stable internal world model that continuously integrates 3D perception, ego states, and language commands. This allows the model to generate not just context-aware, but temporally coherent plans. The model’s effectiveness is demonstrated by its state-of-the-art performance across both open-loop trajectory planning and complex closed-loop driving tasks. Its success highlights a unique capability: bridging perception with action through a persistent, evolving strategy, rather than disconnected, stimulus-response mappings.

they are fused with historical state information and natural language instructions, allowing a set of lightweight, trainable adapters to steer the powerful, pre-trained VLM core towards generating coherent and actionable trajectories.

4.1. Temporal Coherence Module

To maintain a coherent world model over time, the agent must solve the critical challenge of tracking object states despite ego-motion and occlusions. Our TCM, detailed in Alg. 1, addresses this through an elegant, three-stage process of geometric propagation, motion-conditioned state refinement, and evidence-based fusion.

Geometric Propagation. The process begins by explicitly compensating for the vehicle’s own movement. His-

torical 3D object queries $Q_{hist,p}$ are geometrically warped into the current frame’s coordinate system using the ego-motion transformation \mathcal{E}_{ego} . This provides a crucial geometric prior, ensuring that the initial state estimate for each object is grounded in physical reality before any feature-level processing occurs.

Motion-Conditioned State Refinement. Simple geometric alignment is insufficient to capture complex object dynamics or perspective shifts. We therefore introduce a novel motion-aware normalization scheme that dynamically refines the propagated state. Instead of using static normalization parameters, we parameterize conditional affine transformation coefficients, α and β , as a function of the full motion context $(E_{ego}, v, \Delta t)$. These coefficients then

Algorithm 1 Temporal Coherence Module

Require: Historical queries $Q_{hist.c}, Q_{hist.p}$; Current queries $Q_{init.c}$; Image features $F_{img.t}$; Ego-motion $E_{ego}, v, \Delta t$.

Ensure: Refined queries for current frame $Q_{refined.t}$.

- 1: \triangleright 1. *Motion-Aware Propagation*
- 2: $Q_{aligned.p} \leftarrow E_{ego} \cdot Q_{hist.p}$ \triangleright Align historical positions
- 3: $\alpha, \beta \leftarrow \text{MotionEncoder}(E_{ego}, v, \Delta t)$ \triangleright Compute affine parameters
- 4: $Q_{pe} \leftarrow \alpha \cdot \text{LN}(\psi(Q_{aligned.p})) + \beta$ \triangleright Modulate positional embeddings
- 5: $Q_m \leftarrow \alpha \cdot \text{LN}(Q_{hist.c}) + \beta$ \triangleright Propagate context queries

- 6: \triangleright 2. *Hybrid Attention and Fusion*
- 7: $Q_{hybrid} \leftarrow \text{Concat}(Q_m, Q_{init.c})$
- 8: $Q'_{hybrid} \leftarrow \text{SelfAttention}(Q_{hybrid})$
- 9: $Q_{refined.t} \leftarrow \text{CrossAttention}(Q'_{hybrid}, F_{img.t} + Q_{pe}, F_{img.t})$
- 10: **return** $Q_{refined.t}$

perform a motion-conditioned modulation on both the positional embeddings Q_{pe} and the propagated context features Q_m . This allows the network to learn a sophisticated, feature-level compensation, for instance, by amplifying features for fast-moving objects or down-weighting features for objects that are likely occluded.

State Reconciliation and Fusion. The refined memory queries, Q_m , now representing a strong temporal prior, are concatenated with new perception queries, $Q_{init.c}$, which represents fresh evidence from the current frame. A self-attention mechanism then performs state reconciliation, allowing the model to weigh historical beliefs against new observations, update existing tracks, and suppress redundant or spurious detections. Finally, these reconciled queries are grounded back into the current visual evidence through cross-attention. By injecting the modulated positional embeddings Q_{pe} into the keys of the image features, we provide explicit spatial guidance, enabling the model to precisely localize and update the state of each object.

This unified design achieves robust object permanence and temporally consistent perception, forming the bedrock upon which the agent’s cognitive inertia is built.

4.2. 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:

$$\mathcal{L}_{pc} = \lambda_c \mathcal{L}_{cls} + \lambda_r \mathcal{L}_{reg} + \lambda_{mc} \mathcal{L}_{mcls} + \lambda_{mr} \mathcal{L}_{mreg}, \quad (1)$$

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:

$$\mathcal{L}_{total} = \mathcal{L}_{pc} + \mathcal{L}_{ce}. \quad (2)$$

5. Experiments

5.1. Implementation Details

For visual feature extraction, we employ EVA-02-L [7] as our vision encoder, which is pretrained through masked image modeling with knowledge distillation from CLIP [31]. This approach ensures strong alignment between visual features and language representations. LLaVA v1.5 [21] 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 [23] 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. More details are provided in the Appendix.

5.2. Closed-Loop Trajectory Planning Task

As shown in Table 2, we evaluate our model on the challenging Bench2Drive benchmark for closed-loop trajectory planning. Our method establishes a new state-of-the-art by a significant margin on the primary metrics, achieving a Driving Score of 78.21 and a Success Rate of 56.93%. These results represent a 22% and a substantial 63% relative improvement over the previous best-performing methods, respectively. This dramatic improvement in long-term planning success underscores the effectiveness of instilling cognitive inertia, as our agent maintains coherent intentions and avoids the decision jitter that plagues prior works. Furthermore, our model remains highly competitive on other key metrics, achieving the second-best open-loop Avg. L2 score (0.63 vs. 0.62) and a comparable Efficiency score

Table 2. Comparison of open-loop and closed-loop performance on the Bench2Drive dataset. The best results are highlighted in **bold**. The up arrow (\uparrow) indicates that higher is better, while the down arrow (\downarrow) indicates that lower is better.

Method	Open-loop Metric	Closed-loop Metric			
	Avg. L2 \downarrow	Driving Score \uparrow	Success Rate(%) \uparrow	Efficiency \uparrow	Comfortness \uparrow
AD-MLP [47]	3.64	18.05	0.00	48.45	22.63
UniAD-Tiny [11]	0.80	40.73	13.18	123.92	47.04
UniAD-Base [11]	0.73	45.81	16.36	129.21	43.58
VAD [15]	0.91	42.35	15.00	157.94	46.01
TCP [43]	1.70	40.70	15.00	54.26	47.80
ThinkTwice [13]	0.95	62.44	31.23	69.33	16.22
DriveAdapter [12]	1.01	64.22	33.08	70.22	16.01
MomAD [37]	0.87	44.54	16.71	170.21	48.63
DriveTransformer [14]	0.62	63.46	35.01	100.64	20.78
ReAL-AD [24]	0.84	41.17	11.36	–	–
CogAD [42]	–	48.30	24.00	142.00	40.37
Ours	0.63	78.21 (22% \uparrow)	56.93 (63% \uparrow)	169.52	20.50

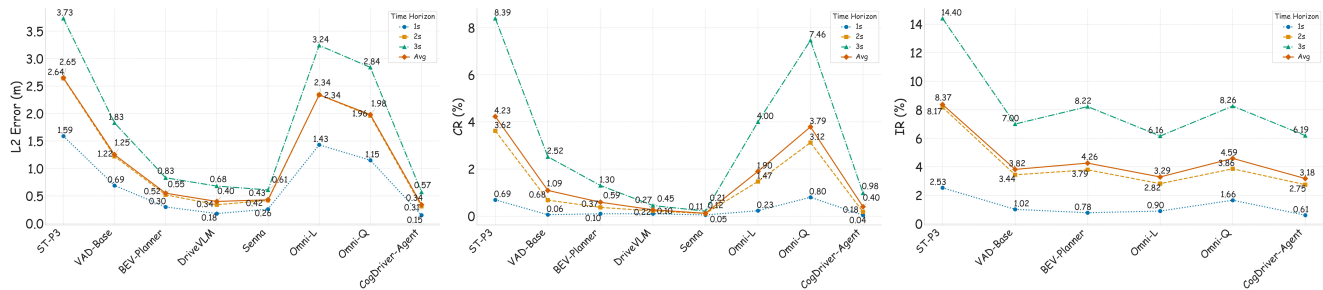


Figure 3. Open-loop planning results on the nuScenes dataset. We compare CogDriver-Agent against several baselines on L2 (left), CR (middle), and IR (right). Performance is reported for 1-3 s horizons and their average. Lower values are better. Our method demonstrates state-of-the-art or highly competitive performance across all metrics.

(169.52 vs. 170.21), demonstrating its all-around strong performance without sacrificing imitation quality or planning efficiency.

5.3. Open-Loop Trajectory Planning Task

As shown in Fig. 3, we compare our proposed CogDriver-Agent against a comprehensive set of state-of-the-art open-loop planning baselines on the nuScenes dataset, including ST-P3 [10], VAD [15], BEV-Planner [20], DriveVLM [39], Senna [16], and OmniDrive (Omni-L, Omni-Q) [40]. The proposed CogDriver-Agent achieves the lowest average L2 distance of 0.63 m, outperforming other baselines such as BEV-Planner with 0.83 m, DriveVLM with 0.40 m, and VAD-Base with 1.25 m. In terms of collision rate (CR), CogDriver-Agent attains an average value of 0.40%, which is slightly higher than Senna’s best result of 0.12% and DriveVLM’s 0.27%, but remains lower than most other baselines. For infraction rate (IR), CogDriver-Agent sets a new state-of-the-art with an average of 3.18%, outperforming all previous methods reporting this metric, in-

cluding BEV-Planner with 4.26%, VAD-Base with 3.82%, and Omni-Q with 4.59%. Across all evaluated horizons, CogDriver-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.

5.4. Driving Visual Question Answering Task

As shown in Table 3, CogDriver-Agent achieves significant performance improvements across all metrics on both CogDriver-nuScenes and CogDriver-Bench2Drive benchmarks compared to state-of-the-art open-source MLLMs. On CogDriver-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 CogDriver-Bench2Drive remains consistently superior, with CogDriver-Agent achieving 9.0%, 53.6%, and 210.8% relative improvements in

Table 3. Comparison of VQA performance conducted on the CogDriver-nuScenes and CogDriver-Bench2Drive datasets. The results show that CogDriver-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. CI-r: CIDEr, BL-1: BLEU-1, BL-4: BLEU-4, ME-R: METEOR, RO-L: ROUGE-L.

Model	CogDriver-nuScenes					CogDriver-Bench2Drive				
	CI-r	BL-1	BL-4	ME-R	RO-L	CI-r	BL-1	BL-4	ME-R	RO-L
Qwen2.5VL 72B [2]	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 [38]	57.10	25.76	4.46	31.81	26.56	78.60	8.21	1.67	83.38	22.60
InternVL3 14B [52]	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 [18]	53.54	4.71	0.47	76.09	15.63	60.08	5.35	0.63	78.53	17.51
CogDriver-Agent	92.39	51.54	14.45	64.45	32.75	95.46	47.25	18.06	58.78	36.99



Figure 4. Visualization of Temporally Coherent Reasoning by CogDriver-Agent. We present two challenging driving scenarios: a left turn in clear conditions (top) and a right lane change in adverse weather (bottom). For each, we visualize the agent’s frame-by-frame narrative predictions. The agent demonstrates cognitive inertia by maintaining a consistent high-level plan (e.g., ‘Left Turn’). Crucially, the underlying rationale is not static; it evolves as the scene unfolds, maturing from reacting to a ‘car ahead’ (Frame 1) to anticipating an ‘upcoming junction’ (Frame 4), proving its capacity for sophisticated, long-term reasoning.

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 CogDriver-Agent’s enhanced multimodal reasoning capabilities but also establish new state-of-the-art performance for VQA tasks in autonomous driving scenarios.

5.5. Ablation Study

5.5.1. Analysis on Different Language Components.

Table 4 ablates the language components of CogDriver-Agent. The full configuration yields the best performance with a 51.54 BLEU-1 score. Individual components provide targeted benefits: environment context boosts language generation by 7.6% in BLEU-1, while dynamic and static object descriptions are critical for safety, achieving the low-

Table 4. Ablation study on the CogDriver-nuScenes dataset. We assess CogDriver-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.

ID	Environment	Dynamic	Static	Reasoning	Action	VQA			Open-Loop		
	Description	Object	Scene			BL-1 ↑	Precision ↑	Recall ↑	L2 ↓	CR ↓	IR ↓
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 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.

TCM	VQA			Open-Loop		
	BL-1 ↑	Precision ↑	Recall ↑	L2 ↓	CR ↓	IR ↓
	52.24	54.84	54.86	0.38	0.44	3.65
✓	51.54	53.32	55.19	0.34	0.40	3.18

est collision rate of 0.37 and intersection rate of 3.08, respectively. Importantly, trajectory prediction remains robust across all configurations with a stable 0.34 L2 error, demonstrating that enhanced reasoning does not compromise planning accuracy.

5.5.2. 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.

5.5.3. Inference Efficiency and Throughput

We evaluate the inference efficiency of CogDriver-Agent on a single NVIDIA A800 GPU. As shown in Table 6, compared to the Qwen2.5VL 32B baseline, CogDriver-Agent achieves significantly higher throughput, reaching 3410.81 input tokens/s and 391.43 output tokens/s. This high processing speed confirms that our architecture satisfies the strict latency constraints required for real-time autonomous driving.

Table 6. Speed comparison of VLMs.

Model	pixels = 256 × 256, tokens = 300	
	Speed input (tokens/s)	Speed output (tokens/s)
Qwen2.5VL 32B	970.94	168.23
CogDriver-Agent	3410.81	391.43

5.6. Qualitative Results

Fig. 4 provides a qualitative analysis of our agent’s reasoning process, demonstrating both decision coherence and causal cognitive coherence.

In the ‘Left Turn’ scenario (top), the agent exhibits strong decision coherence by consistently maintaining the high-level plan ‘Left Turn’ across all frames, avoiding the decision jitter common in reactive models. More critically, the figure reveals the agent’s causal cognitive coherence: the underlying rationale is not static but evolves as the scene unfolds. The reasoning matures from a reactive concern for the ‘car ahead’ (Frame 1) to a strategic, forward-looking plan based on the ‘upcoming junction’ (Frame 4). This demonstrates a crucial capability: the agent’s internal ‘story’ of why it is acting becomes more sophisticated as it gathers more evidence. A similar evolution from immediate hazard avoidance to strategic path planning is observed in the adverse weather scenario (bottom). This stands in stark contrast to stateless models, whose reasoning would remain fixated on the most immediate visual cues without forming a coherent, evolving causal narrative. Additional results are provided in the Appendix.

6. Conclusion

This work presents CogDriver, 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 CogDriver-Agent model that distills human-like priors and causal reasoning, we enable more context-aware and ex-

plainable 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

- [1] Hidehisa Arai, Keita Miwa, Kento Sasaki, Kohei Watanabe, Yu Yamaguchi, Shunsuke Aoki, and Issei Yamamoto. CoVLA: Comprehensive Vision-Language-Action Dataset for Autonomous Driving. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1933–1943. IEEE, 2025. 3, 4
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025. 7
- [3] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with LLMs: Fusing Object-Level Vector Modality for Explainable Autonomous Driving. In *2024 IEEE International Conference on Robotics and Automation*, pages 14093–14100. IEEE, 2024. 3
- [4] Shaoyu Chen, Bo Jiang, Hao Gao, Bencheng Liao, Qing Xu, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. VADv2: End-to-End Vectorized Autonomous Driving via Probabilistic Planning. *arXiv preprint arXiv:2402.13243*, 2024. 1
- [5] Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Luc Van Gool, and Marie Francine Moens. 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*, pages 2088–2098, 2019. 4
- [6] Thierry Deruyttere, Dusan Grujicic, Matthew B Blaschko, and Marie-Francine Moens. Talk2Car: Predicting Physical Trajectories for Natural Language Commands. *Ieee Access*, 10:123809–123834, 2022. 4
- [7] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 149:105171, 2024. 5
- [8] Zhiqi Ge, Hongzhe Huang, Mingze Zhou, Juncheng Li, Guoming Wang, Siliang Tang, and Yueting Zhuang. WorldGPT: Empowering LLM as Multimodal World Model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7346–7355, 2024. 1
- [9] Wencheng Han, Dongqian Guo, Cheng-Zhong Xu, and Jianbing Shen. DME-Driver: Integrating Human Decision Logic and 3D Scene Perception in Autonomous Driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3347–3355, 2025. 3
- [10] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. ST-P3: End-to-End Vision-Based Autonomous Driving via Spatial-Temporal Feature Learning. In *European Conference on Computer Vision*, pages 533–549. Springer, 2022. 6
- [11] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-Oriented Autonomous Driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17853–17862, 2023. 6
- [12] Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, and Hongyang Li. DriveAdapter: Breaking the Coupling Barrier of Perception and Planning in End-to-End Autonomous Driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7953–7963, 2023. 6
- [13] Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, and Hongyang Li. Think Twice Before Driving: Towards Scalable Decoders for End-to-End Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21983–21994, 2023. 6
- [14] Xiaosong Jia, Junqi You, Zhiyuan Zhang, and Junchi Yan. DriveTransformer: Unified Transformer for Scalable End-to-End Autonomous Driving. *arXiv preprint arXiv:2503.07656*, 2025. 6
- [15] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. VAD: Vectorized Scene Representation for Efficient Autonomous Driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8350, 2023. 6
- [16] Bo Jiang, Shaoyu Chen, Bencheng Liao, Xingyu Zhang, Wei Yin, Qian Zhang, Chang Huang, Wenyu Liu, and Xinggang Wang. Senna: Bridging Large Vision-Language Models and End-to-End Autonomous Driving. *arXiv preprint arXiv:2410.22313*, 2024. 6
- [17] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual Explanations for Self-Driving Vehicles. In *Proceedings of the European Conference on Computer Vision*, pages 563–578, 2018. 3
- [18] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. LLaVA-NeXT-Interleave: Tackling Multi-image, Video, and 3D in Large Multimodal Models. *arXiv preprint arXiv:2407.07895*, 2024. 7
- [19] Hongyang Li, Yang Li, Huijie Wang, Jia Zeng, Huilin Xu, Pinlong Cai, Li Chen, Junchi Yan, Feng Xu, Lu Xiong, et al. Open-sourced Data Ecosystem in Autonomous Driving: the Present and Future. *arXiv preprint arXiv:2312.03408*, 2023. 3
- [20] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahao Li, Jan Kautz, Tong Lu, and Jose M Alvarez. 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*, pages 14864–14873, 2024. 6
- [21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 5
- [22] Mingyu Liu, Ekim Yurtsever, Jonathan Fossaert, Xingcheng Zhou, Walter Zimmer, Yuning Cui, Bare Luka Zagar, and Alois C Knoll. A Survey on Autonomous Driving Datasets: Statistics, Annotation Quality, and a Future Outlook. *IEEE Transactions on Intelligent Vehicles*, 2024. 1, 3
- [23] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

- [24] Yuhang Lu, Jiadong Tu, Yuexin Ma, and Xinge Zhu. ReAL-AD: Towards Human-Like Reasoning in End-to-End Autonomous Driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27783–27793, 2025. 6
- [25] Srikanth Malla, Chiho Choi, Isht Dwivedi, Joon Hee Choi, and Jiachen Li. DRAMA: Joint Risk Localization and Captioning in Driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1043–1052, 2023. 3, 4
- [26] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. GPT-Driver: Learning to Drive with GPT. *arXiv preprint arXiv:2310.01415*, 2023. 3
- [27] Jiageng Mao, Junjie Ye, Yuxi Qian, Marco Pavone, and Yue Wang. A Language Agent for Autonomous Driving. *arXiv preprint arXiv:2311.10813*, 2023. 3
- [28] AM Marcu, L Chen, J Hünermann, A Karnsund, B Hanotte, P Chidananda, S Nair, V Badrinarayanan, A Kendall, J Shotton, et al. LingoQA: Visual Question Answering for Autonomous Driving. *arXiv preprint arXiv:2312.14115*, 2023. 3
- [29] Ming Nie, Renyuan Peng, Chunwei Wang, Xinyue Cai, Jianhua Han, Hang Xu, and Li Zhang. Reason2Drive: Towards Interpretable and Chain-Based Reasoning for Autonomous Driving. In *European Conference on Computer Vision*, pages 292–308. Springer, 2024. 4
- [30] Chenbin Pan, Burhaneddin Yaman, Tommaso Nesti, Abhirup Mallik, Alessandro G Allievi, Senem Velipasalar, and Liu Ren. VLP: Vision Language Planning for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14760–14769, 2024. 3
- [31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, pages 8748–8763. PmlR, 2021. 1, 5
- [32] Katrin Renz, Long Chen, Ana-Maria Marcu, Jan Hünermann, Benoit Hanotte, Alice Karnsund, Jamie Shotton, Elahe Arani, and Oleg Sinavski. CarLLaVA: Vision language models for camera-only closed-loop driving. *arXiv preprint arXiv:2406.10165*, 2024. 3
- [33] Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel Kochenderfer, Chiho Choi, and Behzad Dariush. Rank2Tell: A Multimodal Driving Dataset for Joint Importance Ranking and Reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7513–7522, 2024. 4
- [34] Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S Refaat, Rami Al-Rfou, and Benjamin Sapp. MotionLM: Multi-Agent Motion Forecasting as Language Modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8579–8590, 2023. 3
- [35] Hao Shao, Yuxuan Hu, Letian Wang, Guanglu Song, Steven L Waslander, Yu Liu, and Hongsheng Li. LM-Drive: Closed-Loop End-to-End Driving with Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15120–15130, 2024. 3
- [36] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beilwenger, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with Graph Visual Question Answering. In *European Conference on Computer Vision*, pages 256–274. Springer, 2024. 3, 4
- [37] Ziyang Song, Caiyan Jia, Lin Liu, Hongyu Pan, Yongchang Zhang, Junming Wang, Xingyu Zhang, Shaoqing Xu, Lei Yang, and Yadan Luo. Don’t Shake the Wheel: Momentum-Aware Planning in End-to-End Autonomous Driving. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22432–22441, 2025. 6
- [38] Qwen Team. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*, 2024. 7
- [39] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. *arXiv preprint arXiv:2402.12289*, 2024. 6
- [40] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. OmniDrive: A Holistic Vision-Language Dataset for Autonomous Driving with Counterfactual Reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22442–22452, 2025. 4, 6, 10
- [41] Wenhao Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, et al. DriveMLM: Aligning Multi-Modal Large Language Models with Behavioral Planning States for Autonomous Driving. *arXiv preprint arXiv:2312.09245*, 2023. 3
- [42] Zhennan Wang, Jianing Teng, Canqun Xiang, Kangliang Chen, Xing Pan, Lu Deng, and Weihao Gu. CogAD: Cognitive-Hierarchy Guided End-to-End Autonomous Driving. *arXiv preprint arXiv:2505.21581*, 2025. 6
- [43] Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided Control Prediction for End-to-end Autonomous Driving: A Simple yet Strong Baseline. *Advances in Neural Information Processing Systems*, 35:6119–6132, 2022. 6
- [44] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. DriveGPT4: Interpretable End-to-End Autonomous Driving Via Large Language Model. *IEEE Robotics and Automation Letters*, 2024. 3
- [45] Ilker Yildirim and LA Paul. From task structures to world models: what do llms know? *Trends in Cognitive Sciences*, 28(5):404–415, 2024. 1
- [46] Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. RAG-Driver: Generalisable Driving Explanations with Retrieval-Augmented In-Context Learning in Multi-Modal Large Language Model. *arXiv preprint arXiv:2402.10828*, 2024. 3
- [47] Jiang-Tian Zhai, Ze Feng, Jinhao Du, Yongqiang Mao, Jiang-Jiang Liu, Zichang Tan, Yifu Zhang, Xiaoqing Ye,

- and Jingdong Wang. Rethinking the Open-Loop Evaluation of End-to-End Autonomous Driving in nuScenes. *arXiv preprint arXiv:2305.10430*, 2023. 6
- [48] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 1
- [49] Changyuan Zhao, Ruichen Zhang, Jiacheng Wang, Gaosheng Zhao, Dusit Niyato, Geng Sun, Shiwen Mao, and Dong In Kim. World Models for Cognitive Agents: Transforming Edge Intelligence in Future Networks. *arXiv preprint arXiv:2506.00417*, 2025. 1
- [50] Xingcheng Zhou and Alois C Knoll. GPT-4V as Traffic Assistant: An In-depth Look at Vision Language Model on Complex Traffic Events. *arXiv preprint arXiv:2402.02205*, 2024. 1
- [51] Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C Knoll. Vision Language Models in Autonomous Driving: A Survey and Outlook. *IEEE Transactions on Intelligent Vehicles*, 2024. 1
- [52] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*, 2025. 7

CogDriver: Integrating Cognitive Inertia for Temporally Coherent Planning in Autonomous Driving

Supplementary Material

7. CogDriver-Data

7.1. Human-Centric Reasoning Priors

To further enhance the interpretability and human-likeness of the generated captions, we introduce human-centric reasoning priors. This component incorporates explicit prompts centered around driving rules, visual reasoning principles, and scenario-specific heuristics, as shown in Fig. 5b. 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.

7.2. Scene-Aware Grounding

The foundation of our annotation pipeline is Scene-Aware Grounding, a process designed to combat model hallucination by providing an objective, factual basis for all subsequent reasoning. To achieve this, we systematically extract verifiable data directly from the driving logs, rather than relying on a model to infer basic facts. As illustrated in Fig. 5c, this includes precise map context, vehicle dynamics, trajectory profiles, and the states of surrounding agents and traffic signals. This objective information is then structured into a dedicated component of our prompt, forcing the MLLM to condition its output on this unimpeachable ground truth. This methodology ensures that all generated narratives are firmly rooted in the physical reality of the scene, drastically reducing factual errors and providing a reliable foundation for VLA research.

7.3. Narrative Generation with Cognitive Inertia

The generation of narratives with cognitive inertia hinges on the model’s ability to ground static principles in dynamic, unfolding events. To enable this, we condition our powerful MVST-MLLM (72B) simultaneously on the structured prompts from our scene-aware grounding and expert driving principles, and the multi-view, multi-frame video itself, as highlighted in Fig. 5.

Crucially, unlike single-frame captioning solutions that can only produce reactive descriptions, our method operates on temporal windows (5 frames). This is the key mechanism that allows the model to perceive not just a state, but a

state transition. It learns to connect the “before” and “after,” enabling it to generate explanations grounded in causality and persistent intent, the very essence of cognitive inertia. For each 30-second driving scene, overlapping temporal windows comprehensively sample the spatiotemporal dynamics. The resulting output is therefore not a mere caption, but a narrative fragment that explains the current action by linking it to past observations and future goals. Leveraging large-scale distributed inference, our pipeline efficiently produces over 227k such temporally grounded and causally enriched narratives daily, establishing a new standard for high-quality VLA resources.

7.4. Data Analysis

7.4.1. Environment Analysis.

Due to the large data volume and high diversity of Bench2Drive, we primarily showcase analysis on the Bench2Drive dataset. Fig. 6 and Fig. 7 illustrate the structural diversity and complexity of road layouts within Bench2Drive. As shown in Fig. 6, 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. 7 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 multi-faceted 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. 8 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, cap-



System

You are a helpful assistant.

User

You are a certified professional driving instructor and currently demonstrating driving on all kinds of roads.

- The traffic participant information is represented in the 2D top-down view under the corresponding lane, with attributes like class, orientation, position (x, y) in meters, and velocity (vx, vy) in meters per second.
- Definitions for x and y directions: If $x > 0$, it means that the object is in front of your own car, and vice versa. If $y > 0$, it means that the object is to the left of your own car, and vice versa.
- The lane centerline for vehicle navigation, is defined by an ordered list of discrete points. These points are sequentially connected between straight-line segments to form the lane geometry. with-flow: A lane designated for vehicles that are traveling in the same direction as you. opposite-flow: A lane designated for vehicles that are traveling in the opposite direction as you. allowing from right to left driving: A lane that is perpendicular to your current direction, allowing left movement. allowing from left to right driving: A lane that is perpendicular to your current direction, allowing right movement.
- Traffic sign attributes are assigned to the corresponding lane.
- The description and action briefly describe the scene you are in.
- Your status is represented as "your own car" under the corresponding centerline.
- There is one expert trajectory. The expert trajectory meets safety requirements.

Task: As an autonomous vehicle scene understanding assistant, you are given multi-view videos from six perspectives: CAM_FRONT_LEFT , CAM_FRONT , CAM_FRONT_RIGHT , CAM_BACK_LEFT , CAM_BACK , CAM_BACK_RIGHT . Please analyze the following aspects. All outputs must be in English only and do not use any Chinese.

1. "Environment description": Describe overall environmental information, such as weather, road conditions, lighting, etc. Determine whether the vehicle is at or approaching an intersection. Identify the type of road (highway, city road, expressway, rural road, etc.). Recognize road markings, the number of lanes, intersection markings, and traffic signs, etc.
2. "Dynamic object analysis": Extract and summarize the dynamic object information provided by the scene information that may affect the driving, such as vehicles, pedestrians, cyclists, etc.
3. "Static scene analysis": Extract and summarize the static scenarios provided by the scene information that may affect driving, such as lanes, crosswalk, traffic lights(red/green/yellow), stop lines, traffic signs, cones, etc.
4. "Causal reasoning": A brief rationale discussing the cause for the current action/decision.

(a) Example of input and system prompt.

The rules for the dynamic and static obstacles of interest are as follows:

- Dynamic obstacles: 1. Within 20 meters in front of and behind the ego vehicle, and 15 meters laterally. 2. Within 50 meters in front and 30 meters behind the ego vehicle, and 30 meters laterally, and for each lane only keep the closest obstacle to the ego vehicle.
- Traffic signs: Within 30 meters in front of the ego vehicle and 30 meters laterally.
- Traffic lights: Within 50 meters in front of the ego vehicle and 30 meters laterally.

You should obey the following rules when answering questions:

- You should drive on the right side of the road according to the driving rules.
- Do not mention proprietary terms like 'expert trajectory/decision', 'example trajectory/decision'. Always answer as if you are directly in the driving scene.
- Pay attention to the subordinate relationship between objects and lanes, and list them out. For example, a truck traveling ahead is driving on a left turning lane.
- When answering the questions, add numerical information based on the given scene description and action (e.g., objects near your path, you may collide with, traffic signs and signals, etc.), always with their locations, states and belonging lanes if possible.
- When answering the questions, avoid overreporting the threatened traffic elements. For example, objects (stationary or not moving in the same direction) behind you, objects in the distance (having large location values and don't hinder your path), even in the same lane, usually pose no threat. You could say "In this scenario, only xxx is far behind the vehicle/there are no objects around us, which won't affect the driving behavior."

Visual Principle:

- Only describe objects that are clearly visible in the actual image. If an object is not visually present—due to occlusion, being out of the camera's field of view, or insufficient visibility—do not mention or infer its existence in your response. All answers must be strictly based on what can be directly seen and verified in the image.

Analytical Focus:

- Go beyond simply listing or copying items from the scene.
- For each section, think about how specific aspects could influence driving safety, possible decision-making, and what causal impact these elements have on driver choices.
- Highlight not just the current state but also potential risks or changes in the environment, and proactive strategies to maintain safety.

Generalizability:

- Express your analysis in a broadly applicable manner—don't repeat or copy example wording or logic.
- Relate your reasoning to possible variations in similar scenes.
- In each field, provide reasoning that connects scene elements to the decisions or precautions a safe driver would make.
- The output must be in English, no Chinese characters are allowed.

(b) Example of human-centric rules.

```

- The following is the scene information that needs attention:
Weather: clear.
Traffic Light Existing: False.
├── (road_id:4, lane_id:1), your current straight lane [(+3.8, +0.2), (+8.8, +0.4), (+13.8, +0.5), (+18.8, +0.6), (+23.8, +0.8), (+28.8, +0.9),
(+38.8, +1.2), (+43.8, +1.3), (+48.8, +1.4)]
├── your own car
├── pedestrian in the front location: (+7.0, -1.2), velocity: (+0.3, +2.0).
├── car in the back location: (-17.7, -0.4), velocity: (+3.9, +0.1).
└── (road_id:4, lane_id:-1), opposite-flow, straight lane [(+47.9, +5.4), (+42.9, +5.3), (+37.9, +5.1), (+27.9, +4.9), (+22.9, +4.7), (+17.9,
+4.6), (+12.9, +4.5), (+2.9, +4.2)]

- Ego vehicle planning trajectory:
Ego vehicle speed: 4.99
Expert decision: Moving Slowly, Decelerate, Lane Keeping
Expert trajectory: [PT, (+1.32, +0.03), (+2.18, -0.24), (+1.26, +0.39), (+2.29, +0.33), (+3.22, -0.64), (+2.41, +0.27)].
Objects near your path: pedestrian at (+7.0, -1.2).

```

(c) Example of attention and ego states.

Figure 5. Prompt template and input structure for data annotation.

turing complex urban scenes where pedestrian interactions can occur. Overall, these results highlight the dataset’s rich diversity, spanning from dense multi-vehicle interactions to sparse pedestrian scenarios, while also capturing the realistic distribution of safety-critical events, which are infrequent and typically occur at greater distances from the ego vehicle.

7.4.2. Action Analysis.

Fig. 9 and Fig. 10 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.

7.4.3. Causal Reasoning Analysis

Fig. 11 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.

7.4.4. Case Study.

Fig. 12 and Fig. 13 from the CogDriver-Bench2Drive dataset showcase the annotation model’s scene understanding capability and ability to reason effectively in complex and challenging driving scenarios. Fig. 12 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. 13 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.

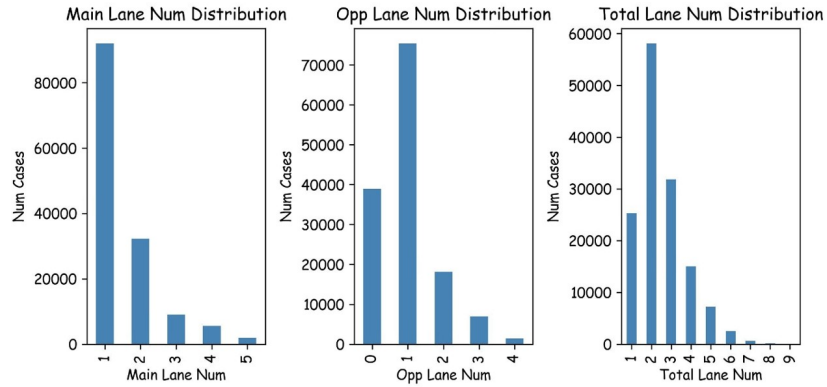


Figure 6. Lane number distributions in the Bench2Drive dataset. The left plot shows the distribution of main (ego-moving direction) lane numbers, with the majority of cases having one or two lanes in the same direction. The middle plot displays the distribution of opposite-direction lanes, where most 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.

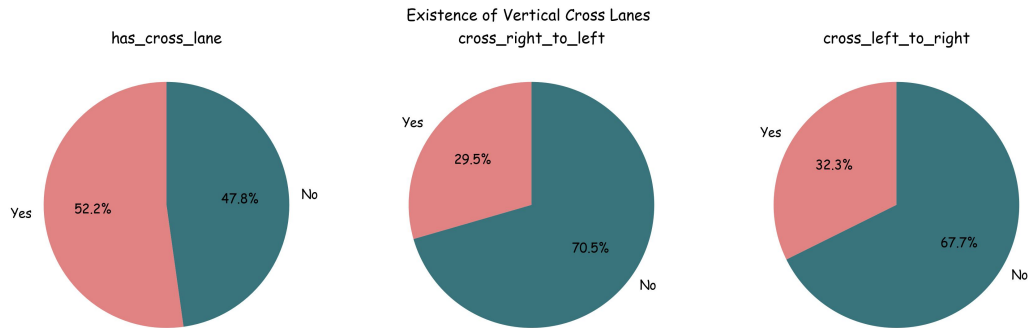


Figure 7. Distribution of cross lanes in the Bench2Drive dataset. The left pie chart shows the proportion of samples with cross lanes (has_cross_lane). The middle and right pie charts illustrate the existence of vertical cross lanes from right to left (cross_right_to_left) and from left to right (cross_left_to_right).

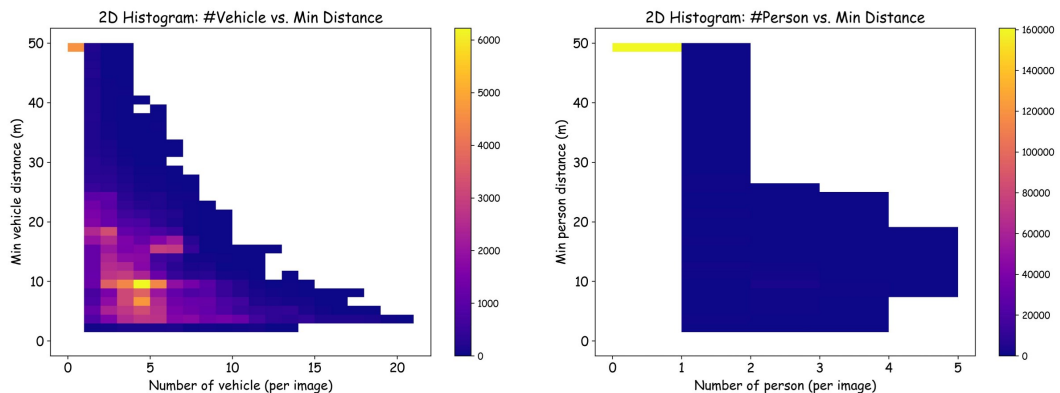


Figure 8. 2D histograms visualizing the relationship between the number of objects and their minimum distance to the ego vehicle in each image. The left plot shows the distribution for vehicles, indicating that as the number of vehicles in the scene increases, the minimum distance to the nearest vehicle 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.

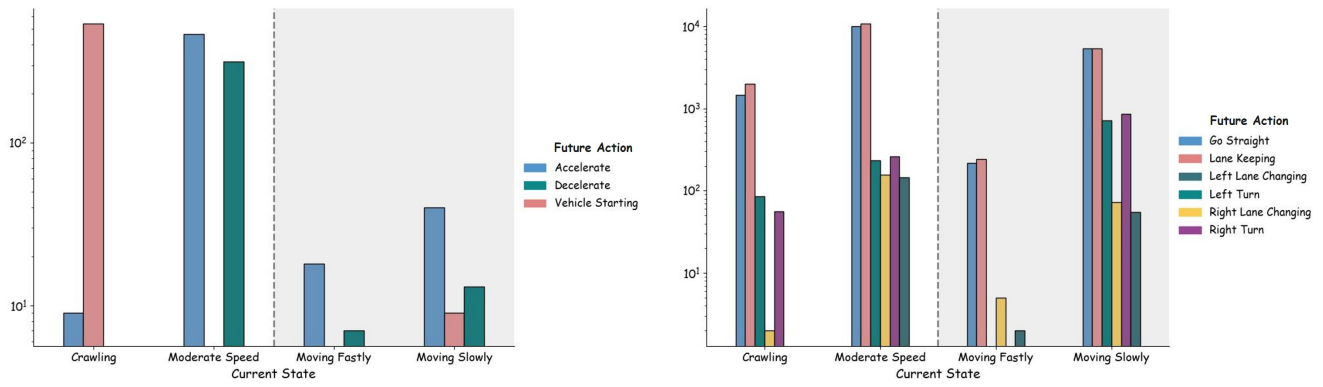


Figure 9. Ego vehicle's future action at different current states on the nuScenes dataset.

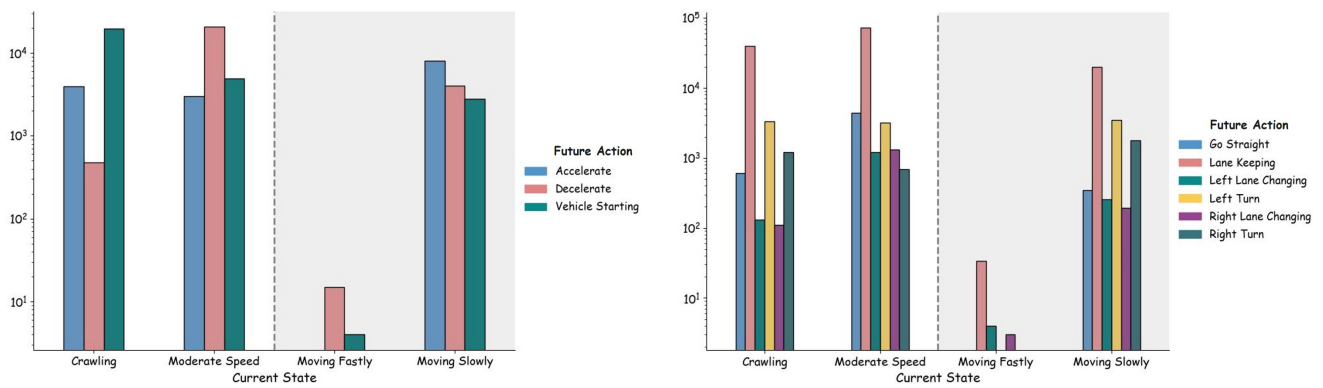


Figure 10. Ego vehicle's future action at different current states on the Bench2Drive dataset.



Figure 11. Word cloud of causal reasoning annotations on the nuScenes (left) and Bench2Drive (right) 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."

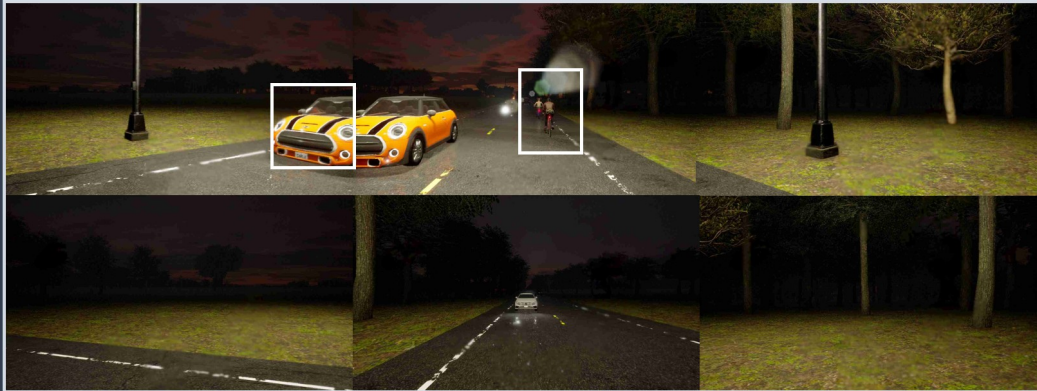
Action

Expert decision: Crawling, Vehicle Starting, Lane Keeping.

Trajectory

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

Figure 12. A dynamic object crossing example from the CogDriver-Bench2Drive dataset is illustrated. The white box represents the dynamic objects around the ego vehicle.



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 13. A hazard at side lane two ways example from the CogDriver-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 14. A stop VQA results of CogDriver-Agent on the nuScenes dataset.

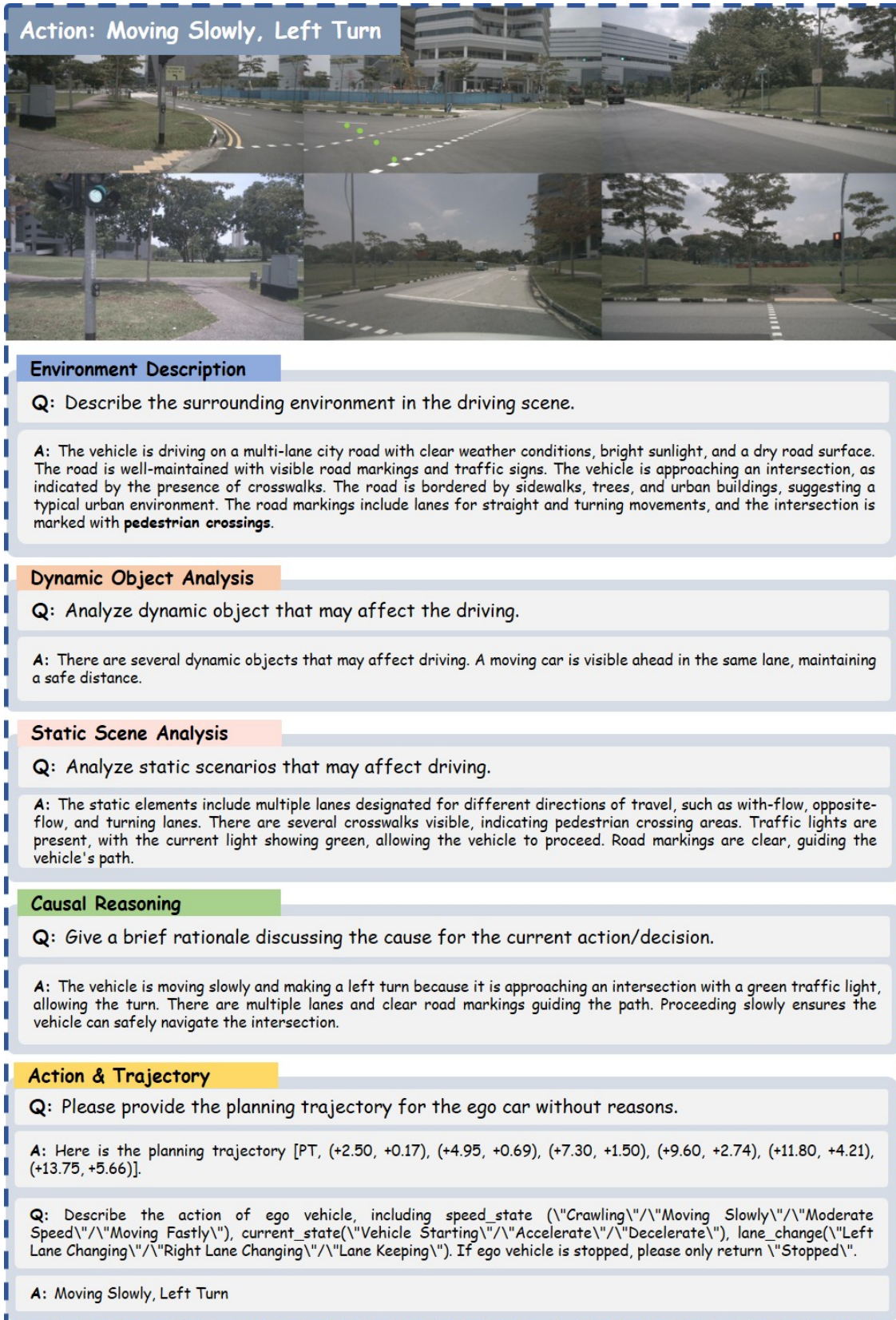


Figure 15. A left turn VQA results of CogDriver-Agent on the nuScenes dataset. The green refers to the predicted trajectory.

8. CogDriver-Agent

The training of CogDriver-Agent follows a two-stage curriculum designed to first align modalities and then fine-tune for the end-to-end driving task. This strategy effectively leverages the frozen LLM core while ensuring stable convergence.

Vision-Language Alignment. The initial stage trains the vision encoder and a set of lightweight adapters to produce meaningful features for the frozen LLM. Using our CogDriver-Data, the model learns to generate the ground-truth narrative explanation from multi-view video clips and a structured prompt, the template for which is detailed in Fig. 16.

End-to-End VLA Fine-tuning. In the second stage, we fine-tune the agent for the full vision-language-action task. The vision encoder and adapters remain trainable, while the LLM core stays frozen. The model is optimized to jointly predict both the narrative and the future trajectory waypoints.

```

System
You are a helpful assistant.
User
You are a certified professional driving instructor. Please
answer the question based on the following principles:
- The traffic participant information need to be
represented in the 2D top-down view under the corresponding
lane, with attributes like class, orientation, position (x,
y) in meters, and velocity (vx, vy) in meters per second.
- Definitions for origin, x and y directions: The origin is
located at the center of the ego vehicle's front axle. If x
> 0, it means that the object is in front of your own car,
and vice versa. If y > 0, it means that the object is to
the left of your own car, and vice versa. The unit is
(meters, m/s).

Task: As an autonomous vehicle scene understanding
assistant, you are given multi-view videos from six
perspectives: CAM_FRONT_LEFT , CAM_FRONT , CAM_FRONT_RIGHT ,
CAM_BACK_LEFT , CAM_BACK , CAM_BACK_RIGHT .
<Question>

Please strictly follow the structured JSON format for
output:
<Answer>

```

Figure 16. Prompt template for CogDriver-Agent.

9. Experiments

9.1. Datasets and Evaluation Metrics

The proposed CogDriver-Data encompasses open-loop planning, closed-loop planning, and VQA tasks. We evaluate CogDriver-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). To assess performance in more realistic, interactive scenarios, we evaluate closed-loop planning on the

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
CogDriver-Agent	86.67	36.40	7.52	47.58	27.87

Table 7. Performance comparison on the OmniDrive dataset [40]. CogDriver-Agent outperforms prior works in all metrics.

Bench2Drive dataset using Driving Score, Success Rate, Efficiency, and Comfortness. 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), ME-TEOR (ME-R), ROUGE-L (RO-L), Precision, and Recall, ensuring a comprehensive analysis of language understanding and multimodal alignment.

9.2. Ablation Study

9.2.1. Generalizability on OmniDrive Dataset.

Table 7 presents a comprehensive evaluation of CogDriver-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 CogDriver-Agent’s architectural innovations deliver superior results regardless of model scale, establishing new state-of-the-art performance for autonomous driving applications.

9.3. Qualitative Results

Fig. 14 and Fig. 15 showcase the VQA results of CogDriver-Agent on the nuScenes dataset for two distinct urban intersection scenarios. In Fig. 14, 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. 15 presents a left turn scenario where the agent recognizes a green traffic light and



Figure 17. Qualitative comparison of decision coherence.

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.

Fig. 17 visualizes the causal alignment between decisions and trajectories. While the baseline (Omni-Q) suffers from severe instability, erratically flipping between “Go Straight” and “Right Turn”, CogDriver-Agent demonstrates robust Cognitive Inertia. Our CogDriver-Agent maintains a consistent, long-term driving intent that aligns perfectly with the Ground Truth, proving its ability to generate stable, causally grounded plans despite visual fluctuations.