

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.

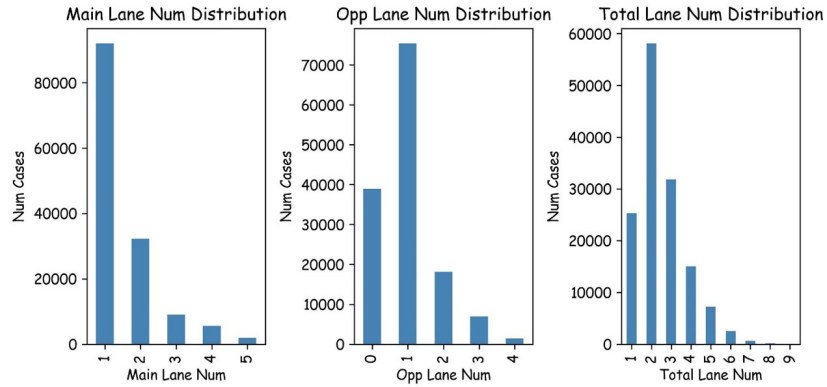


Figure 6. Lane number distributions in the Bench2Drive dataset. The left plot shows the distribution of main (ego-moving direction) lane numbers, with the majority of cases having one or two lanes in the same direction. The middle plot displays the distribution of opposite-direction lanes, where most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

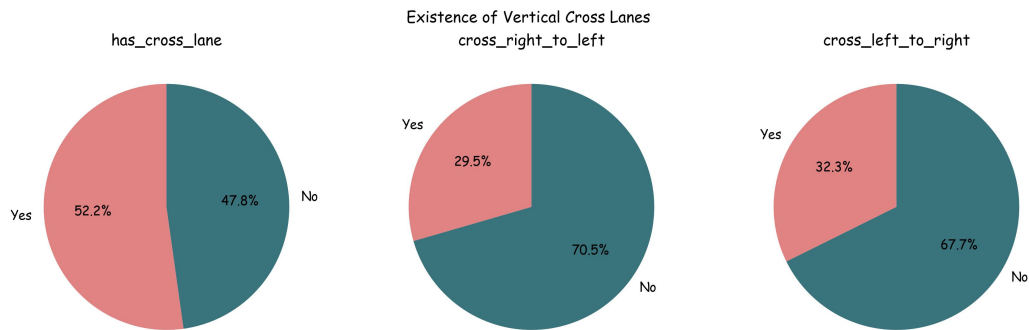


Figure 7. Distribution of cross lanes in the Bench2Drive dataset. The left pie chart shows the proportion of samples with cross lanes (has_cross_lane). The middle and right pie charts illustrate the existence of vertical cross lanes from right to left (cross_right_to_left) and from left to right (cross_left_to_right).

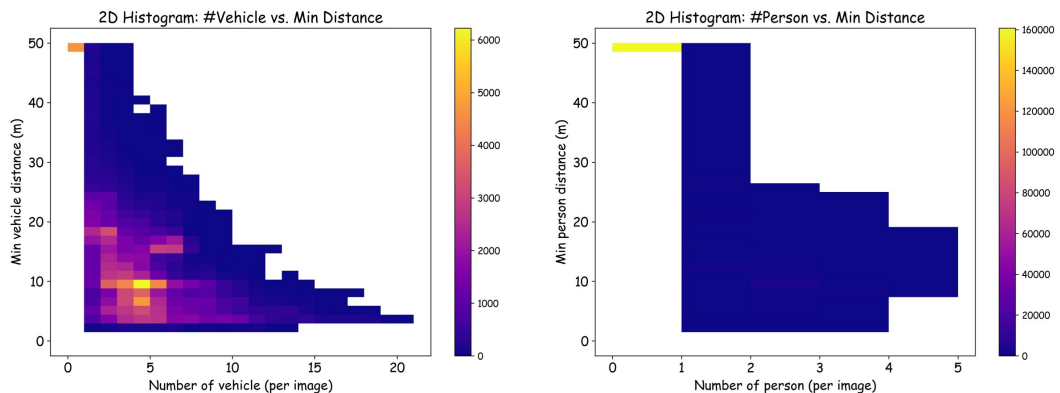


Figure 8. 2D histograms visualizing the relationship between the number of objects and their minimum distance to the ego vehicle in each image. The left plot shows the distribution for vehicles, indicating that as the number of vehicles in the scene increases, the minimum distance to the nearest vehicle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



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."

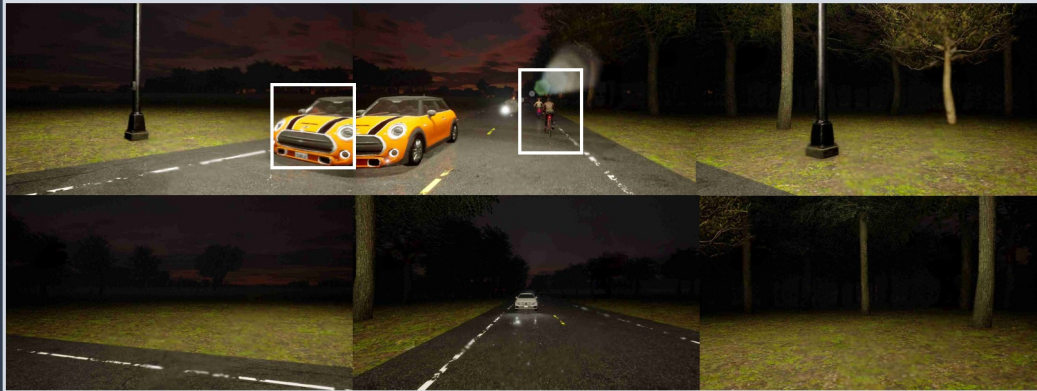
Action

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.

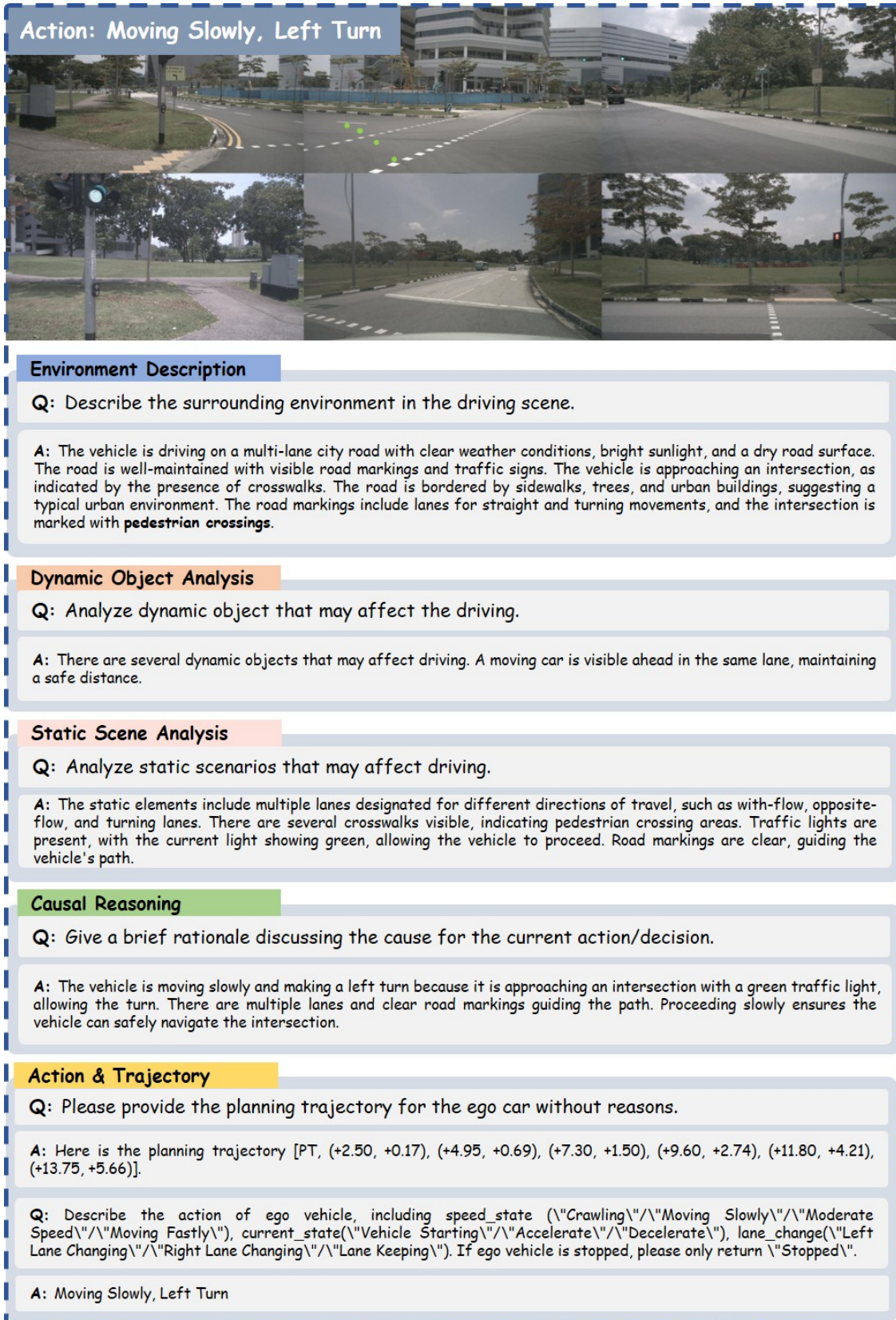


Figure 15. A left turn VQA results of CogDriver-Agent on the nuScenes dataset. The green refers to the predicted trajectory.

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), METEOR (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.