More Than Meets the Eye? Uncovering the Reasoning-Planning Disconnect in Training Vision-Language Driving Models

Xurui Song[1,2*] **Shuo Huai**[1,*] **Jingjing Jiang**[1] **Jiayi Kong**[1,2] **Jun Luo**[1,†]

[1]College of Computing and Data Science, Nanyang Technological University, Singapore
[2]S-Lab, Nanyang Technological University, Singapore

Abstract

Vision-Language Model (VLM) driving agents promise explainable end-to-end autonomy by first producing natural-language reasoning and then predicting trajectory planning. However, whether planning is **causally** driven by this reasoning remains a critical but unverified assumption. To investigate this, we build DriveMind, a large-scale driving Visual Question Answering corpus with plan-aligned Chain-of-Thought (CoT), automatically generated from nuPlan. Our data generation process converts sensors and annotations into structured inputs and, crucially, separates priors from to-be-reasoned signals, enabling clean information ablations. Using DriveMind, we train representative VLM agents with Supervised Fine-Tuning and Group Relative Policy Optimization and evaluate them with nuPlan's metrics. Our results, unfortunately, indicate a consistent **causal disconnect** in reasoning-planning: removing ego/navigation priors causes large drops in planning scores, whereas removing CoT produces only minor changes. Attention analysis further shows that planning primarily focuses on priors rather than the CoT. Based on this evidence, we propose the Reasoning-Planning Decoupling Hypothesis, positing that the training-yielded reasoning is an ancillary byproduct rather than a causal mediator. To enable efficient diagnosis, we introduce a novel, training-free probe that measures an agent's reliance on priors by evaluating its planning robustness against minor input perturbations. In summary, we provide the community with a new dataset and a diagnostic tool to evaluate the **causal fidelity** of future models.

1 Introduction

End-to-end autonomous driving learns planning directly from sensor data and has attracted sustained attention in both academia and industry [7, 6, 18, 19]. Recent studies explore Vision-Language Model (VLM) driving agents using visual question answering (VQA) that combines the reasoning capability of large language models (LLMs) with visual perception to approximate human driving [37, 42]. Chain-of-Thought (CoT) [36] has been shown to enhance reasoning in LLMs [12], and it is increasingly adopted in VLM driving agents to make the sequence of perception, analysis, and decision explicit [31, 33, 35]. The intention is to strengthen planning by improving controllability and interpretability. In this paradigm, the model first articulates a CoT reasoning for the current driving, followed by the final planning trajectory. Consequently, planning is taken for granted as causally driven through the preceding reasoning.

*Equal contribution.
†Corresponding author.
Code and dataset: `https://github.com/SONG0257/DriveMind`

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Despite rapid progress, whether planning in current VLM driving agents is causally mediated by their reasoning remains insufficiently verified. Existing works [39, 34] primarily report trajectory quality and rule compliance, which assess how well the planning appears, but not which information pathway produce it. As a result, strong scores cannot be taken as evidence that reasoning contributes causally to planning. Under these conditions, shortcut learning [14, 41] is a central risk: a model can obtain high planning scores by exploiting biased or spurious priors, such as ego state and history, rather than by using reasoning to construct the planning. Consequently, reasoning and planning may become decoupled, raising a critical safety concern: even when a plan appears safe, its underlying safety-relevant logic cannot be reliably verified, raising the risk that such outcomes are merely probabilistic coincidences rather than causally grounded decisions suitable for deployment.

To systematically investigate this causal link, a high-quality VQA dataset with plan-aligned CoT is necessary. However, existing datasets fall short of this need. Many real-world datasets [31, 30] use the nuScenes [5] benchmark as their foundation. While nuScenes provides high-fidelity sensor data, it inherently lacks the fine-grained semantic annotations, such as traffic light states, or complex lane topology, which are essential for deep reasoning. To circumvent the semantic limitations inherent to nuScenes, some researches [35] have turned to simulation platforms such as CARLA [10]. While highly controllable, these simulators suffer from a pronounced sim-to-real gap [9]: their trajectories follow idealized dynamics and fail to capture the nuanced nature of real-world human driving.

To bridge this gap, we introduce DriveMind, a high-quality dataset curated to facilitate the causal analysis of VLM-based driving agents. Built upon the nuPlan [21] benchmark that combines the authenticity of large-scale, real-world driving data with the rich, vectorized semantic context necessary for complex reasoning, DriveMind covers approximately 50K samples spanning 61 driving scenarios, providing broad and diverse coverage for analysis. Another core contribution of DriveMind is the modular organization of the dataset's multi-modal inputs, structuring elements like visual data, ego state, and navigation into distinct modules. This design is critical for conducting controlled ablation studies. By selectively withholding specific information modalities, we can systematically dissect the information flow within a VLM agent and robustly attribute the final planning decisions to either high-level reasoning or low-level shortcut signals.

Using our DriveMind dataset, we train and evaluate representative VLM-based driving agents [3, 25, 34]. We uncover a striking result: an agent trained solely on textual priors, with no visual input and no CoT reasoning, achieves planning scores that match or even exceed those of a fully multimodal counterpart. This reliance on shortcuts is so entrenched that even applying an advanced policy alignment method from reinforcement learning (RL), Group Relative Policy Optimization (GRPO), fails to substantively restore the causal link from reasoning to planning. These findings motivate our Reasoning-Planning Decoupling Hypothesis: the planning module from an agent's output predominantly relies on textual priors (i.e., ego state, history) as shortcuts, largely ignoring the visual context (i.e., surroundings, traffic signals) and the CoT reasoning. To substantiate this hypothesis, we introduce a sequence-level attention analysis, which demonstrates the dominance of prior and ego-state tokens over visual and CoT tokens.

Finally, we propose a generalizable, training-free diagnostic method to distinguish between genuine, CoT-grounded planning and shortcut learning, with the aim of establishing a plug-and-play standard for model evaluation. Grounded in the principles of causal intervention, our method acts as a "causal probe" by applying minor, semantically plausible perturbations to the textual priors (e.g., slight variations in ego states or historical positions). A planner that truly grounds decisions in visual evidence and CoT should be stable under such perturbations, whereas a shortcut-reliant planner will show brittle sensitivity to the exact prior pattern. Therefore, a disproportionately large degradation in planning scores following perturbation indicates a high degree of shortcut learning. Conversely, a robust planning performance signifies that the agent's decisions are properly grounded in a holistic understanding of the driving scene. This method provides a simple yet powerful tool for assessing the causal fidelity of VLM agents. In summary, the main contributions of our paper are as follows:

- We propose and validate the Reasoning-Planning Decoupling Hypothesis, which posits that current training paradigms are insufficient to forge a causal link between reasoning and planning, leading agents to instead learn shortcuts from textual priors.
- We introduce DriveMind, a large-scale nuPlan-based dataset with plan-aligned CoT and a modular design, enabling causal analysis and systematic ablation studies of VLM driving agents.

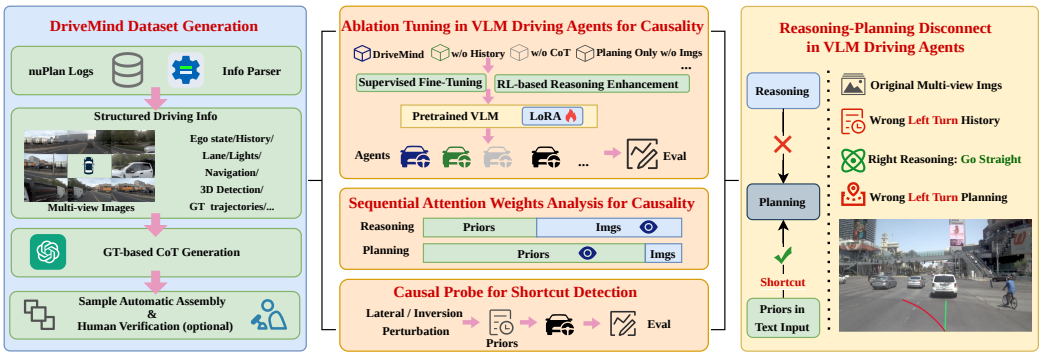


Figure 1: An overview of our framework for investigating the reasoning-planning disconnect in VLM agents. (Left) The DriveMind dataset generation pipeline. (Middle) Our methodology for causal validation. (Right) An illustration of our core finding, where planning follows a textual prior shortcut, bypassing the correct CoT reasoning.

- We introduce the Causal Probe, a novel, training-free diagnostic method that detects shortcut reliance, providing a new tool for evaluating the causal robustness of driving agents.

2 Related Work

End-to-End Autonomous Driving. Unlike hierarchical driving models [2], end-to-end autonomous driving aims to learn directly from sensor inputs and output planning trajectories, thereby avoiding error accumulation and facilitating optimization. Most existing end-to-end autonomous driving models are based on vector representation learning [6]. Frameworks such as UniAD [18] and VAD [19] transform perception information into Bird Eye View (BEV) feature vectors for subsequent trajectory planning. ViDAR [40] and UAD [15] proposed self-supervised pre-training algorithms, enhancing the scalability of vector representation-based driving models. More recently, the remarkable success of LLM [4, 1, 20] and their multimodal extension, VLMs [11, 25, 43], has introduced a new frontier. Researchers have begun to leverage the powerful feature extraction capabilities of VLMs to enhance the vectorized paradigm. For instance, models like VLP [28] and DiMA [16] distill prior knowledge from pre-trained VLMs into the driving model to improve planning performance. However, both the purely vectorized models and these early VLM-enhanced frameworks fundamentally rely on implicit feature transformations for decision-making. They lack an explicit, human-understandable reasoning process. Moreover, recent studies reveal that such models are prone to shortcut learning on ego status [23] and spurious correlations [29, 32]. VLMs, generating explicit reasoning before planning, are thus regarded as promising for countering this by improving interpretability and controllability [38].

Driving Agent with Reasoning. To enhance the interpretability of VLMs and adapt their capabilities for the driving domain, recent research has focused on optimizing models toward reasoning-based planning. Early approaches often employed multi-turn dialogue paradigms to elicit complex reasoning. For instance, DriveLM [31] introduced a graph-based VQA dataset that requires the model to first perceive and predict before a human manually selects key context for planning. Similarly, DriveVLM [33] proposed a hierarchical thinking dataset structured around scenario description, analysis, and planning. However, this multi-turn paradigm incurs significant data annotation and inference overhead, posing challenges to scalability and real-time application. To improve efficiency, subsequent work has shifted towards single-turn dialogue incorporating CoT. DriveCoT [35], for example, explicitly integrates reasoning and planning into a single model response, but its reliance on the CARLA [10] simulation environment raises key questions about sim-to-real transferability. Omnidrive [34] trains an agent for implicit reasoning via multi-task VQA. Meanwhile, Omnidrive recognizes the model's dependency on priors like ego state and attempts to mitigate this using counterfactual reasoning. However, as our subsequent experiments will reveal, this method has limited effectiveness in addressing shortcut learning from textual priors. Furthermore, its dataset is built upon nuScenes, inheriting the limitations of semantic richness required for deep causal analysis.

3 Methodology

This section outlines our methodology, which includes four parts. First, we introduce the DriveMind dataset and its generation pipeline. Second, we describe the ablation training of VLM driving agents on DriveMind. Third, we present our sequence-level attention analysis for interpreting information flow. Finally, we introduce our perturbation-based causal probe for assessing shortcut learning.

3.1 DriveMind Dataset Generation Pipeline

Existing datasets either lack the complexity of the real world due to being simulation-based or lack the rich semantic information required to support deep reasoning, and they are not designed for causal analysis between reasoning and planning for VLM-based driving agents. Therefore, the design of DriveMind adheres to four core principles, including a real-world foundation, rich semantic context, plan-aligned CoT, plus a modular structure for causal analysis. As a large-scale planning benchmark released by Motional, nuPlan uniquely combines the high fidelity of real-world data with the rich semantics (e.g., traffic light status, lane topology) necessary for deep reasoning. Based on nuPlan, we convert the raw log data into the final samples of the DriveMind dataset through an automated data generation pipeline , as illustrated in Figure 1. Our data generation pipeline consists of three main stages, beginning with scene parsing and feature extraction, followed by GPT-4.1-based CoT generation, and concluding with automatic sample construction with optional human verification.

The first stage involves developing a parser to structure all key information from each nuPlan log. We extract ego-vehicle priors such as current and historical states, as well as the global navigation goal. For visual information, all camera images are first rescaled and then stitched into a three-row image grid corresponding to the front, side, and rear view groups to provide a spatially coherent input. For lane information, we encode it by identifying the ego lane and its neighboring lanes, along with their widths and directions relative to the vehicle heading, reflecting road topology and drivable regions. Additionally, we include traffic signals by extracting the current state of relevant lights and the position of their corresponding stop lines. Finally, we extract features of scene objects within a 20 m radius, categorizing valid targets into four groups: vehicles, pedestrians, bicycles, and other obstacles (e.g., traffic cones). These targets are localized using 3D annotations from nuPlan ground truth, and we then employ GPT-4.1 to identify visual characteristics (e.g., 'a white SUV') of each object from their cropped images. This process enhances the agent's perception by creating an explicit link between each visual object and its corresponding textual description.

In the second stage, we use GPT-4.1 to assist in generating high-quality CoT. Recognizing that current LLMs face challenges in precise 3D spatial reasoning from images alone [13], we ground the reasoning process by providing GPT-4.1 with structured scene ground truth. We design a comprehensive multimodal prompt that includes the preprocessed visual input, the structured scene ground truth from the first stage, and the expert's future planning trajectory. GPT-4.1 is tasked with explaining the causal logic behind the expert trajectory. The resulting CoT is organized into a structured text with three parts, which are scene ground truth, a causal analysis, and a macro decision.

Finally, we construct the final samples and perform human verification. To ensure the quality of the automatically generated CoT, a random 10% ($\sim$5K) of the samples are manually reviewed by human experts, with only 2 out of 5K reviewed samples (0.04%) required regeneration, confirming a high degree of logical correctness of our generation pipeline. The model is provided with two inputs, the visual inputs containing the preprocessed images and the textual priors containing ego state, history, and navigation information. The model is trained to produce two targets, the complete, human-verified ground truth CoT, and the ground truth planning from the nuPlan human-expert driving logs. The details of our reasoning generation process and an example including the complete input and target labels for a training sample are provided in Appendix A.4.

Our training set is constructed from 61 driving scenarios sourced from the nuPlan training split. To mitigate the inherent long-tail distribution of this real-world data, we employ square-root weighted stratified sampling. For a target dataset size of $M = 50,000$ across $K = 61$ scenarios, the number of samples n_i drawn from the i-th scenario (which originally contains N_i samples) is calculated as:

$$n_i = M \cdot \frac{\sqrt{N_i}}{\sum_{j=1}^{K} \sqrt{N_j}}, \tag{1}$$

By this sampling, our DriveMind dataset maintains a high degree of complexity with 33.9% hard scenarios (e.g., turnings and dense pedestrian interactions), 39.2% general scenarios (e.g., slow down

Table 1: Overview of ablation settings for tuning analysis. By systematically removing priors, ablation studies are conducted to examine the impact of CoT and driving priors on VLM-based driving agents. The 'decomposed CoT' means using multi sub-tasks for training. GRPO-based experiments confirm that the reasoning-planning disconnect effect is intrinsic rather than a superficial artifact.

Agent	Base Model	Tuning	Vision	CoT	Priors
Supervised Fine-Tuning (SFT) Experiments					
Base	Qwen2.5-vl-7b	—	✓	—	—
CoT	Qwen2.5-vl-7b	SFT	✓	✓	All
Plan	Qwen2.5-vl-7b	SFT	✓	✗	All
Plan_NoV	Qwen2.5-vl-7b	SFT	✗	✗	All
CoT_NoHis	Qwen2.5-vl-7b	SFT	✓	✓	w/o History
CoT_NoHis_Ego	Qwen2.5-vl-7b	SFT	✓	✓	w/o His, Ego
CoT_NoPri	Qwen2.5-vl-7b	SFT	✓	✓	None
CoT_L	llava1.6-mistral-7b	SFT	✓	✓	All
Plan_L_NoV	llava1.6-mistral-7b	SFT	✗	✗	All
Omnidrive	Omni	SFT	✓	Decomposed	All
Omnidrive_NoPri	Omni	SFT	✓	Decomposed	None
GRPO Policy Alignment Experiments					
CoT_grpo	Qwen2.5-vl-7b	SFT+GRPO	✓	✓	All
Base_grpo	Qwen2.5-vl-7b	GRPO	✓	Self-learned	All
Base_grpo_NoPri	Qwen2.5-vl-7b	GRPO	✓	Self-learned	None

at lights), while simple scenarios (e.g., stationary) comprise only 26.9%. This ensures that the model is trained primarily on scenarios requiring active reasoning rather than simple path following or being stationary. Besides the dense semantic context derived from the real-world nuPlan logs and the high-quality ground truth based CoT, the critical feature of our dataset is its modular information structure, which is designed specifically for causal analysis. By explicitly separating different priors, reasoning, and planning within a coherent text, researchers can perform precise information ablation. This allows for a reliable determination of whether an agent's planning stems from its autonomous visual perception and high-level reasoning, or from a reliance on "shortcut learning" from the provided textual priors. The detailed distribution of all 61 scenario types and comparison with other datasets are provided in Appendix A.5 and A.6.

3.2 Tuning VLMs for Causal Analysis

Since general VLMs are not pre-trained for driving planning, a crucial first step is to fine-tune them into proficient driving agents. Our primary objective, however, is to use the fine-tuning process itself as our object of study. We investigate how current training paradigms such as supervised fine-tuning (SFT) shape the causal structure of an agent's learned behaviors, for testing our Reasoning-Planning Decoupling Hypothesis. To this end, we conduct controlled experiments on a primary representative VLM, Qwen2.5-vl [3], and verify the universality of our findings on a second VLM, Llava-1.6 [24]. Crucially, our setting is made comprehensive by also including Omnidrive [34], which employs counterfactual reasoning in an attempt to mitigate prior dependency. By analyzing it, we can test our hypothesis against existing state-of-the-art solutions. This three-pronged approach, which covers in-depth analysis, architectural generalization, and state-of-the-art comparison, provides a robust and sufficient foundation for verifying our claims.

We first employ SFT to create a set of agents under various information ablation conditions. At this stage, we utilized the DriveMind dataset and its variants shown in Table 1, where key input modalities, such as visual information, textual priors, or CoT are systematically removed. We fine-tune an agent with LoRA [17] using each ablation dataset $D = \{((I_i, T_i), \mathcal{Y}_i)\}_{i=1}^{M}$, where M is the total number of samples in the set. For each sample i, I_i represents the visual input, T_i is the textual prompt, and $\mathcal{Y}_i = (y_{i,1}, y_{i,2}, \ldots, y_{i,L})$ is the corresponding target output sequence of length L. With the pre-trained VLM parameters frozen as θ and the trainable matrices as ϕ, we perform SFT, and the cross-entropy loss is used to compare two probability distributions for each output token:

$$\mathcal{L}_{\text{SFT}}(\phi) = -\frac{1}{M} \sum_{i=1}^{M} \sum_{j=1}^{L} \log P(y_{i,j} \mid I_i, T_i, y_{i,<j}; \theta, \phi). \tag{2}$$

To test the robustness of this disconnect finding, we then incorporate an RL-based policy alignment stage using GRPO, which explicitly enhances an agent's reasoning capabilities by rewarding the CoT process. We further sample $1,000$ new VQA samples from nuPlan, maintaining the same scenario distribution as DriveMind, and then do GRPO on both the SFT-trained model and the base model

without SFT, with the latter serving as a baseline. Let D_{grpo} denote the dataset and π_θ the policy, GRPO aims to maximize the following objective [8]:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{\substack{(I,T)\sim D_grpo \\ \{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}}} \left[\frac{1}{G}\sum_{i=1}^{G} (\mathcal{A}_i\, \tilde{w}_i - \beta\, \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})) \right]. \tag{3}$$

This objective means that for each pair of inputs, G outputs are sampled from an older policy $\pi_{\theta_{\text{old}}}$, then each output is evaluated its normalized advantage $\mathcal{A}_i$ that measures its quality relative to the group's average. The $\tilde{w}_i$ is a weight to stabilizes updates by preventing overly large policy ratios and $\mathbb{D}_{\text{KL}}$ is a KL-divergence penalty to regularize the current policy π_θ to stay close to the base policy π_{ref} with the tunable hyperparameter β. The model is encouraged to improve planning performance via enhanced reasoning by three rewards in $\mathcal{A}$: a location reward, a velocity reward, and a format reward. The first two rewards measure planning accuracy while the format reward incentivizes reasoning ability. For detailed explanations of GRPO and our reward functions, please refer to Appendix A.7. This stage serves as a critical test to demonstrate that the observed disconnect is not an artifact of the SFT process, but a persistent characteristic that remains even after a direct intervention designed to strengthen the reasoning-planning link.

3.3 Diagnosing with Sequence-level Attention Analysis

To investigate our hypothesis and gain insight into the VLM's internal mechanisms, we employ an interpretability framework called Sequence-level Attention Analysis. It aims to analyze patterns of macro-level information flow between reasoning and planning in the agent's output. When the model outputs trajectory tokens, we measure how much attention it pays to the preceding CoT tokens versus other contextual information (e.g., ego-state, history). A high degree of attention on the reasoning sequence would indicate a strong correlational link, suggesting the plan is at least conditioned on the reasoning. Conversely, low attention would provide evidence in line with our decoupling hypothesis.

We formalize this by analyzing the attention weights of the VLM, which has L transformer layers and H attention heads per layer. Let $a_{j\to i}^{(l,h)}$ denote the raw attention weight that the token at position i (the query) assigns to the token at position j (the key) in layer l and head h. Our analysis then proceeds in two steps. First, we calculate a sequence-aggregated attention score, $\bar{a}_{j\to S_t}^{(l)}$. This score represents the total attention that a target sequence S_t (e.g., the planning) directs to a single preceding token j, averaged over all H attention heads in a given layer l:

$$\bar{a}_{j\to S_t}^{(l)} = \frac{1}{H}\sum_{i\in S_t}\sum_{h=1}^{H} a_{j\to i}^{(l,h)}. \tag{4}$$

Second, using this aggregated score, we compute our primary metric: the proportional attention score, $G_{S_M\to S_t}^{(l)}$. This value calculates the proportion of the target sequence's (S_t) total attention that is allocated to a specific source sequence S_M (e.g., the CoT reasoning). Given a set of M_{total} preceding source sequences $\{S_1, S_2, \ldots, S_{M_{\text{total}}}\}$, this attention proportion is given by:

$$G_{S_M\to S_t}^{(l)} = \left(\sum_{j\in S_M} \bar{a}_{j\to S_t}^{(l)}\right) / \left(\sum_{m=1}^{M_{\text{total}}}\sum_{j\in S_m} \bar{a}_{j\to S_t}^{(l)}\right). \tag{5}$$

This metric, which sums to 1 across all source sequences, provides a direct and quantitative measure of information dependency. By comparing the attention proportion directed towards the reasoning sequence versus other contextual sequences, we can systematically verify whether an agent's planning is effectively grounded in its own reasoning.

3.4 Causal Probe for VLM Driving Agents

Standard performance metrics (e.g., planning scores) are insufficient to distinguish between planning that originates from genuine reasoning and that relies on shortcut learning. To efficiently and deeply probe an agent's decision-making logic and diagnose its degree of reliance on shortcuts, we introduce a novel, training-free causal probe. The probe operates on the principle of robustness: a planning decision effectively driven by high-level reasoning from the visual scene should be largely invariant to minor, semantically plausible perturbations in its textual priors. Conversely, an excessive reaction to such minor changes indicates a brittle reliance on those priors. Our framework introduces two distinct perturbation methods to test this principle.

The first method is a quantitative test we term *lateral offset perturbation*. In this probe, we apply a small lateral offset, δ, to the agent's ego-velocity. A robust agent, truly reasoning based on the visual scene and CoT, should be able to ignore or correct for a small offset in its ego velocity as its visual input clearly reflects the vehicle's true position and heading in the world. In contrast, an agent merely pattern-matching the textual priors would be misguided. Let $\mathbf{p}_i \in \mathbb{R}^2$ denote the predicted endpoint coordinates of the original input for the i-th sample, and $\tilde{\mathbf{p}}_i$ represent the endpoint under perturbation, we then quantify the impact by measuring the Mean Final Deviation (MFD):

$$\text{MFD} = \frac{1}{n}\sum_{i=1}^{n}\|\mathbf{p}_i - \tilde{\mathbf{p}}_i\|_2, \tag{6}$$

where n is the number of tested samples. We also evaluate the Relative Intervention Degree (RID) that is defined as the ratio of the difference between a sample's output before and after perturbation to the theoretical error that would occur if the model perfectly followed the perturbation (i.e., $\delta \cdot v_i \cdot T$):

$$\text{RID} = \frac{1}{n}\sum_{i=1}^{n}\frac{\|\mathbf{p}_i - \tilde{\mathbf{p}}_i\|_2}{\delta \cdot v_i \cdot T}, \tag{7}$$

where v_i is the ego-velocity of the i-th sample and T is the prediction horizon. This normalized RID quantitatively measures the extent to which a model relies on textual priors, providing a scale-independent, robust metric that allows for cross-scenario comparisons.

The second method, *lateral direction inversion*, is a qualitative probe. In this probe, we keep the ego-vehicle's current state (ego-state) unchanged but invert the lateral component of its historical trajectory (e.g., mirroring a left-side merging history to a right-side one). A positive diagnosis for shortcut learning occurs when this inversion causes a directional change in the planning (e.g., a right turn), while the generated CoT that is primarily conditioned on the unchanged visual scene still presents correct arguments for a left turn. This exposes a stark contradiction between the agent's stated reasoning and its actual planning logic.

By combining the quantitative measurement and the qualitative diagnosis, we provide a systematic, training-free method to identify the presence and severity of shortcut learning in VLM-based driving agents. Together with our DriveMind dataset, they form a reusable evaluation framework.

4 Experiment Results and Analysis

4.1 Experiment Setup

Implementation Details: We fine-tune the agents listed in Table 1, with the specific training parameters detailed in Appendix A.9. We also verify our hypothesis on the large commercial model GPT-4o in a zero-shot manner.

Evaluation Protocol: For evaluation, we use the official nuPlan Challenge test set. To ensure comprehensive coverage, we randomly sample 200 scenarios for each of the 14 distinct types defined in the benchmark. We conduct both open-loop and closed-loop (non-reactive) evaluations for every agent to holistically assess the performance. More details are provided in Appendix A.10.

4.2 Tuning Ablation Analysis Results

Our primary finding, detailed in Table 2, is the VLM agent's profound reliance on textual priors over CoT reasoning for planning. The *CoT* agent with explicit reasoning and *OmniDrive*, trained using decomposed CoT, achieve performance nearly indistinguishable from the *Plan* agent that produces no reasoning output. This comparison strongly indicates that the articulated reasoning process offers no discernible benefit to the final planning outcome. The most compelling evidence for disconnect is a paradoxical result: the *Plan_NoV* agent, effectively 'driving blind' without any visual input, performs on par with the *CoT* agent. This is a powerful demonstration of shortcut learning. Since the priors contain no information about environmental conditions, it reveals the model has learned to even bypass scene understanding by merely extrapolating from its current state and history. Conversely, the removal of textual priors results in a catastrophic performance collapse, as evidenced by the sharp decline across all metrics for *CoT_NoPri* and *Omnidrive_NoPri*. The consistency of these findings across diverse models, including Qwen-VL, Llava-1.6, and Omnidrive, provides robust evidence for our Reasoning-Planning Decoupling Hypothesis.

To further subject our hypothesis to a rigorous stress test, we employ GRPO, a powerful policy alignment tool. By rewarding high-quality planning outcomes GRPO has the potential to optimize

Table 2: Main ablation results on the nuPlan test set at 1s, 2s, and 3s horizons

	Open-Loop Score ↑			Closed-Loop Score ↑			Avg. ADE / FDE (m) ↓			Collision Ratio ↓		
Driving Agent	**1s**	**2s**	**3s**	**1s**	**2s**	**3s**	**1s**	**2s**	**3s**	**1s**	**2s**	**3s**
Base	36.52	33.62	31.50	59.80	42.48	36.99	2.67/4.45	4.80/8.81	7.02/13.75	7.84%	9.16%	9.87%
CoT	98.91	96.67	92.56	97.46	96.42	92.38	0.03/0.08	0.14/0.39	0.33/1.02	0.00%	0.29%	2.15%
Omnidrive	98.92	96.85	92.96	97.49	96.35	92.47	0.03/0.08	0.13/0.38	0.31/0.99	0.00%	0.31%	2.03%
Plan	98.89	96.64	92.58	97.38	96.49	92.60	0.04/0.09	0.14/0.41	0.33/1.02	0.00%	0.36%	2.55%
Plan_NoV	98.95	96.84	92.92	97.46	96.37	92.54	0.03/0.08	0.13/0.38	0.32/0.98	0.00%	0.36%	2.45%
CoT_NoHis	97.44	89.30	82.82	96.71	94.17	87.93	0.11/0.27	0.38/0.98	0.78/2.08	0.12%	1.62%	5.09%
CoT_His_Ego	74.51	61.65	57.70	90.39	87.23	81.32	0.69/1.25	1.35/2.60	2.04/4.05	0.24%	1.80%	5.27%
CoT_NoPri	72.40	58.58	54.56	89.82	85.87	79.19	0.73/1.34	1.47/2.88	2.26/4.60	0.23%	2.15%	6.58%
Omnidrive_NoPri	75.91	61.34	57.32	89.99	86.52	79.69	0.64/1.17	1.29/2.53	2.00/4.14	0.31%	2.27%	6.54%
CoT_L	98.81	96.49	92.22	97.45	95.82	91.32	0.03/0.08	0.14/0.40	0.34/1.03	0.00%	0.54%	3.05%
Plan_L_NoV	98.91	96.90	93.03	97.54	96.53	92.87	0.03/0.08	0.13/0.36	0.31/0.95	0.00%	0.36%	2.09%
Base_grpo	96.38	85.31	77.84	96.74	92.52	84.51	0.15/0.35	0.47/1.20	0.95/2.50	0.12%	1.74%	6.22%
Base_grpo_NoP	39.93	29.70	27.84	73.88	61.75	55.44	1.80/3.22	3.55/6.88	5.39/10.68	10.59%	15.68%	20.41%
CoT_grpo	98.76	96.35	92.25	97.36	96.26	91.74	0.05/0.10	0.15/0.44	0.36/1.08	0.00%	0.48%	2.45%

Table 3: Planning performance comparison under visual inputs and reasoning perturbations.

Agent	**Open-Loop Score (3s) ↑**	**Close-Loop Score (3s) ↑**
CoT	92.56	92.38
CoT_V_Crop_Noise	91.86	90.79
CoT_V_Replace	92.36	92.27
CoT_R_Replace	92.39	92.17

the model's upstream cognitive chain, making it the most promising method for forging a causal link between reasoning and planning. However, our experiment results directly contradict this optimistic outlook. While the *Base_grpo* agent shows superficial improvements over the *Base* counterpart, this proves to be a "false prosperity" built on shortcuts. Once textual priors are removed, the *Base_grpo_NoPri* agent's performance collapses catastrophically, falling well below even the original *Base* model. This decisively shows that this powerful alignment tool does not guide the model to learn deeper visual reasoning; instead, it rewards model's reliance on textual shortcuts, exacerbating the overfitting problem. Furthermore, the *CoT_grpo* agent's planning capability is even weaker than that of *CoT*, suggesting that GRPO is unable to effectively link reasoning and planning to search better results. Ultimately, this stress test provides strong evidence that the reasoning–planning disconnect is a fundamental characteristic of the current VLM fine-tuning paradigm. Even advanced RL policy designed to boost alignment fails to rectify this disconnect and may even intensify the model's dependence on shortcuts. Appendix A.14 provides additional results supporting our findings.

4.3 Perturbation to Visual Inputs and Generated Reasoning

To further validate our findings regarding the reasoning-planning disconnect, we conduct intervention experiments by perturbing both the visual inputs and the intermediate reasoning traces during inference, evaluating whether the final planning genuinely relies on visual scene understanding or generated reasoning. Specifically, we evaluate three perturbation variants on the *CoT* agent: (1) *CoT_V_Crop_Noise*, which randomly crops the input images and injects Gaussian noise; (2) *CoT_V_Replace*, which randomly substitutes the original visual scene with a contradictory one; and (3) *CoT_R_Replace*, where we replace the macro-driving decision within the generated reasoning trace with an incompatible decision taken from a different driving scenario while keeping the visual input intact, and then allow the agent to continue generating the final planning.

Table 3 reports the planning performance across all perturbation settings. Under visual perturbations, the *CoT* agent exhibits negligible performance changes in both open-loop and closed-loop scores. Together with the findings from *Plan_NoV* that completely strips visual input and serves as the most extreme visual perturbation, these results collectively demonstrate that the planning ability of the trained VLM driving agent is effectively visually inert. Moreover, despite replacing the generated reasoning with a contradictory macro-decision, *CoT_R_Replace* maintains high performance with 92.39 and 92.17 in open-loop and closed-loop scores respectively, suggesting that the final planning

Table 4: Attention distributions across preceding sequences during target sequence generation.

Attention Target	Scenario Reasoning Task			Planning Task		
	Shallow Layer	Middle Layer	Final Layer	Shallow Layer	Middle Layer	Final Layer
Image Tokens (%)	**2.91**	**5.55**	**11.52**	2.47	2.51	1.82
Textual Priors (%)	3.74	5.34	10.36	**11.52**	**17.93**	**26.97**
Generated Reasoning (%)	—	—	—	17.11	21.12	19.37
All Textual Tokens (%)	97.09	94.45	88.48	97.53	97.49	98.18

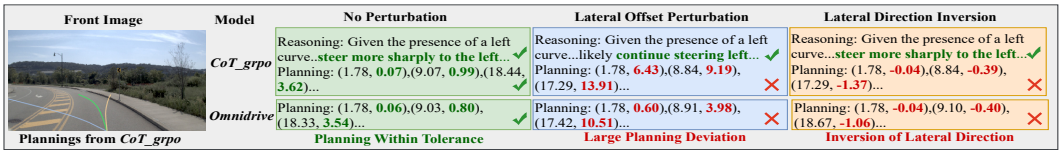


Figure 2: Comparison of an agent's original plan and its trajectories under two perturbations.

process operates largely independently of the generated reasoning. Together, these results provide further support for our Reasoning-Planning Decoupling Hypothesis.

4.4 Sequence-level Attention Analysis Results

To investigate the agent's decision-making process, we analyze the attention flow of the *CoT_grpo* agent. We partition the model's output into two key sequences: the *Reasoning* (CoT) sequence and the *Planning* (trajectory) sequence. We then measure the proportion of attention each sequence directs towards different preceding information sources (Images, Textual Priors, etc.) at shallow, middle, and final layers of the model, allowing us to observe how information dependencies evolve.

As shown in Table 4, when the agent generates the *Reasoning* sequence, its attention to the Image Tokens progressively increases, from 2.91% in the shallow layer to over 11% in the final layer. This indicates that the reasoning process is continuously and increasingly grounded in the visual input. While attention to Textual Priors also grows, it remains secondary to the visual information in deeper layers. This behavior represents a healthy, logical process where the agent builds its reasoning primarily upon what it sees, while integrating priors as necessary context.

In contrast, the attention shifts dramatically during the *Planning* phase. Attention to the model's generated Reasoning first increases and then fades, while attention to the Textual Priors skyrockets from 11.52% to nearly 27% in the final layer. This shows the model discovering the decisive role of priors and becoming highly dependent on them. Simultaneously, attention to the Image Tokens becomes negligible, dropping below 2%, which indicates that the model barely requires visual information for planning. This finding, with the high performance of the *Plan_NoV* agent in Table 2, leads to a clear conclusion: the VLM driving agent relies almost exclusively on textual priors for planning, rendering the reasoning and planning processes almost entirely disconnected.

A consistent observation across both tasks is the VLM's overwhelming preference for attending to textual tokens over image tokens. This highlights an inherent modality bias in current VLMs. This insight suggests that a potential path to mitigating shortcut learning in VLM driving agents is to force the visual understanding capabilities especially when planning.

4.5 Causal Probe Detection Results

To evaluate the diagnostic efficacy of our causal probe, we select two key agents for this analysis: the strongest reasoning agent, *CoT_grpo*, and a state-of-the-art baseline known to include measures against prior dependency, *Omnidrive*. First, we apply the *lateral offset perturbation*. In this test, we add a small lateral velocity, δ, perpendicular to the vehicle's heading, scaled to the ego-vehicle's current speed v_{ego} (i.e., $\delta = 0.1 \cdot v_{\text{ego}}$). This creates a small and semantically plausible perturbation. As shown in Figure 2, both agents exhibit extreme sensitivity to this dynamic perturbation.

For more precise quantitative analysis and to test the model's sensitivity under different perturbation levels, we test on the *CoT_grpo* utilizing a test set of 300 randomly sampled scenarios with varied the perturbation magnitude δ from 0.02 to 0.3, and we report the MFD and RID in Table 5. From the

Table 5: Impact of different perturbation intensities (δ) on *CoT_grpo*. The large MFD and RID demonstrate its strong reliance on text priors.

δ	**0.02**	**0.06**	**0.08**	**0.10**	**0.30**
MFD $\downarrow$	0.55	2.17	3.75	5.89	15.46
RID $\downarrow$	1.00	1.47	2.04	2.17	2.04

results, the model exhibits extreme instability even at low perturbation levels. As δ increases from 0.02 to 0.06, the MFD jumps significantly to 2.17 m. This confirms that the agent is captured by the textual prior almost immediately, lacking the robustness to correct for even subtle semantic conflicts. As δ increases further the MFD continues to grow monotonically, reaching 15.46m at $\delta = 0.3$. The RID initially peaks at 2.17 for $\delta = 0.1$, then plateaus or slightly decreases to 2.04 for $\delta = 0.3$. Actually, this plateau does not indicate robustness as the RID is still larger than 1; rather, it reflects saturation of "wrong outputs" as the agent predicts extreme trajectories beyond its learned manifold, lagging behind increasing perturbations. The disproportionate reactions of the driving agents under small lateral offset perturbation from both the qualitative and quantitative results strongly indicate the agents' heavy reliance on the ego states in text priors for shortcut learning.

Next, we apply the *lateral direction inversion* probe. After inverting the lateral component of the historical trajectory, we observed that both *CoT_grpo* and *Omnidrive* consistently reversed their planned maneuver direction. Most critically, a manual inspection of the *CoT_grpo* agent's outputs after this inversion revealed a direct contradiction in 100% of the tested samples. In these cases, the agent would generate a correct CoT based on the visual scene (e.g., reasoning for a left turn) but then produce a plan that incorrectly followed the perturbed history (e.g., executing a right turn).

This finding provides compelling evidence for our Reasoning-Planning Decoupling Hypothesis. It not only demonstrates the deep-rooted nature of the problem but also validates the efficacy of our causal probes in uncovering these critical failures in reasoning. More examples are shown in Appendix A.12.

4.6 Evaluation on Larger Models

To assess whether our findings generalize to larger VLMs, we apply the same causal probes to the substantially larger GPT-4o in a zero-shot setting. Under perturbation, GPT-4o produces correct scenario reasoning (e.g., "The ego-vehicle is in a lane moving straight"), yet its predicted trajectory still adhered to the falsified textual prior, indicating that the reasoning–planning disconnect persists beyond the 7B model scale. Quantitatively, with a perturbation level of $\delta = 0.1$, GPT-4o yields an MFD of 2.55 m and an RID of 1.35. This zero-shot evaluation provides a necessary assessment of a closed-source model, demonstrating that the disconnect is fundamental to both open-source and proprietary large-scale VLM architectures.

5 Conclusion

In this paper, we have investigated the causal link between reasoning and planning in VLM driving agents. To enable this analysis, we introduce DriveMind, a benchmark designed for causal analysis, and develop diagnostic tools including controlled ablations and perturbation-based probes. Our experiments provide strong evidence for the Reasoning-Planning Decoupling Hypothesis: agents learn to rely on textual priors as shortcuts for planning, while their generated CoT serves as a plausible but non-causal byproduct. This work reveals that perceived interpretability of current agents can be misleading, underscoring the need for causally-aware training paradigms to build truly robust VLM driving agents. Possible future directions and limitations are discussed in Appendix A.1 and A.2.

Acknowledgments

This research is supported by cash and in-kind funding from NTU S-Lab and industry partner(s). The authors declare no competing interests.

References

[1] Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large Language Models for Mathematical Reasoning: Progresses and Challenges. In Neele Falk, Sara Papi, and Mike Zhang, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-srw.17. URL https://aclanthology.org/2024.eacl-srw.17/.

[2] ApolloAuto. Apollo: An open autonomous driving platform. https://github.com/ApolloAuto/apollo, 2025.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL Technical Report, 2025. URL https://arxiv.org/abs/2502.13923.

[4] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4, 2023. URL https://arxiv.org/abs/2303.12712.

[5] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[6] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-End Autonomous Driving: Challenges and Frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10164–10183, 2024. doi: 10.1109/TPAMI.2024.3435937.

[7] commaai. openpilot. https://github.com/commaai/openpilot, 2025.

[8] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, 2025. URL https://arxiv.org/abs/2501.12948.

[9] Elahe Delavari, Feeza Khan Khanzada, and Jaerock Kwon. A Comprehensive Review of Reinforcement Learning for Autonomous Driving in the CARLA Simulator, 2025. URL https://arxiv.org/abs/2509.08221.

[10] Alexey Dosovitskiy, Germán Ros, Felipe Codevilla, Antonio M. López, and Vladlen Koltun. CARLA: An Open Urban Driving Simulator. In *1st Annual Conference on Robot Learning, CoRL 2017, Mountain View, California, USA, November 13-15, 2017, Proceedings*, volume 78 of *Proceedings of Machine Learning Research*, pages 1–16. PMLR, 2017. URL http://proceedings.mlr.press/v78/dosovitskiy17a.html.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[12] Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards Revealing the Mystery behind Chain of Thought: A Theoretical Perspective. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 70757–70798. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/dfc310e81992d2e4cedc09ac47eff13e-Paper-Conference.pdf.

[13] Qiyue Gao, Xinyu Pi, Kevin Liu, Junrong Chen, Ruolan Yang, et al. Do Vision-Language Models Have Internal World Models? Towards an Atomic Evaluation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26170–26195, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1342. URL https://aclanthology.org/2025.findings-acl.1342/.

[14] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://www.nature.com/articles/s42256-020-00257-z.

[15] Mingzhe Guo, Zhipeng Zhang, Yuan He, Ke Wang, and Liping Jing. End-to-End Autonomous Driving without Costly Modularization and 3D Manual Annotation, 2024. URL https://arxiv.org/abs/2406.17680.

[16] Deepti Hegde, Rajeev Yasarla, Hong Cai, Shizhong Han, Apratim Bhattacharyya, Shweta Mahajan, Litian Liu, Risheek Garrepalli, Vishal M. Patel, and Fatih Porikli. Distilling Multi-modal Large Language Models for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27575–27585, June 2025.

[17] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[18] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[19] 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 (ICCV)*, pages 8340–8350, October 2023.

[20] Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A Survey on Large Language Models for Code Generation. *ACM Trans. Softw. Eng. Methodol.*, July 2025. ISSN 1049-331X. doi: 10.1145/3747588. URL https://doi.org/10.1145/3747588.

[21] Napat Karnchanachari, Dimitris Geromichalos, Kok Seang Tan, Nanxiang Li, Christopher Eriksen, Shakiba Yaghoubi, Noushin Mehdipour, Gianmarco Bernasconi, Whye Kit Fong, Yiluan Guo, and Holger Caesar. Towards learning-based planning: The nuPlan benchmark for real-world autonomous driving. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 629–636, 2024. doi: 10.1109/ICRA57147.2024.10610077.

[22] Yue Li, Meng Tian, Dechang Zhu, Jiangtong Zhu, Zhenyu Lin, Zhiwei Xiong, and Xinhai Zhao. Drive-R1: Bridging Reasoning and Planning in VLMs for Autonomous Driving with Reinforcement Learning, 2025. URL https://arxiv.org/abs/2506.18234.

[23] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and Jose M. Alvarez. Is Ego Status All You Need for Open-Loop End-to-End Autonomous Driving? In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14864–14873, 2024. doi: 10.1109/CVPR52733.2024.01408.

[24] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning, 2023.

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=w0H2xGHlkw.

[26] Motional. Motional: nuPlan devkit (Planning Challenge). https://github.com/motional/nuplan-devkit, 2023.

[27] NVIDIA, Yan Wang, Wenjie Luo, Junjie Bai, Yulong Cao, Tong Che, et al. Alpamayo-R1: Bridging Reasoning and Action Prediction for Generalizable Autonomous Driving in the Long Tail, 2026. URL https://arxiv.org/abs/2511.00088.

[28] 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 (CVPR)*, pages 14760–14769, June 2024.

[29] Mozhgan Pourkeshavarz, Junrui Zhang, and Amir Rasouli. CaDeT: A Causal Disentanglement Approach for Robust Trajectory Prediction in Autonomous Driving. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14874–14884, 2024. doi: 10.1109/CVPR52733.2024.01409.

[30] Tianwen Qian, Jingjing Chen, Linhai Zhuo, Yang Jiao, and Yu-Gang Jiang. NuScenes-QA: a multi-modal visual question answering benchmark for autonomous driving scenario. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i5.28253. URL https://doi.org/10.1609/aaai.v38i5.28253.

[31] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with Graph Visual Question Answering. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 256–274, Cham, 2025. Springer Nature Switzerland.

[32] Jiacheng Tang, Zhiyuan Zhou, Zhuolin He, Jia Zhang, Kai Zhang, and Jian Pu. CausalVAD: Deconfounding End-to-End Autonomous Driving via Causal Intervention, 2026. URL https://arxiv.org/abs/2603.18561.

[33] 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. In *8th Annual Conference on Robot Learning*, 2024. URL https://openreview.net/forum?id=928V4Umlys.

[34] 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22442–22452, June 2025.

[35] Tianqi Wang, Enze Xie, Ruihang Chu, Zhenguo Li, and Ping Luo. DriveCoT: Integrating Chain-of-Thought Reasoning with End-to-End Driving, 2024. URL https://arxiv.org/abs/2403.16996.

[36] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf.

[37] Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao MA, Yingxuan Li, Linran XU, Dengke Shang, Zheng Zhu, Shaoyan Sun, Yeqi BAI, Xinyu Cai, Min Dou, Shuanglu Hu, Botian Shi, and Yu Qiao. On the Road with GPT-4V(ision): Explorations of Utilizing Visual-Language Model as Autonomous Driving Agent. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024. URL https://openreview.net/forum?id=2UBexKm8TE.

[38] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are VLMs Ready for Autonomous Driving? An Empirical Study from the Reliability, Data and Metric Perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6585–6597, October 2025.

[39] 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*, 9(10):8186–8193, 2024. doi: 10.1109/LRA.2024.3440097.

[40] Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual Point Cloud Forecasting Enables Scalable Autonomous Driving. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14673–14684, 2024. doi: 10.1109/CVPR52733.2024.01390.

[41] Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. Do LLMs Overcome Shortcut Learning? An Evaluation of Shortcut Challenges in Large Language Models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12188–12200, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.679. URL https://aclanthology.org/2024.emnlp-main.679/.

[42] Jimuyang Zhang, Zanming Huang, Arijit Ray, and Eshed Ohn-Bar. Feedback-Guided Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15000–15011, June 2024.

[43] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-Language Models for Vision Tasks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, 2024. doi: 10.1109/TPAMI.2024.3369699.

[44] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. SWIFT:A Scalable lightWeight Infrastructure for Fine-Tuning, 2024. URL https://arxiv.org/abs/2408.05517.

A Appendix

A.1 Future Work

Our future work will focus on resolving the reasoning-planning disconnect. We plan to pursue two research directions aimed at developing a more robust and causally-faithful VLM driving agent.

Mitigating Modality Bias via Contrastive Pre-Finetuning. Our main results revealed a key finding: an agent with no visual input (*Plan_NoV*) can achieve planning scores nearly identical to a fully multimodal agent. This strongly suggests that the models have learned to almost entirely disregard the visual modality. Therefore, a critical future direction is to introduce a dedicated contrastive pre-finetuning stage designed to make the visual input indispensable. In this stage, we will train the VLM on a curated, driving-related dataset where the textual input is held constant while the visual input varies to produce a different correct output. This process will force the model to extract decisive information from the visual modality, rebalancing its modality dependence and enhancing the salience of the vision encoder before the downstream driving SFT/GRPO begins.

Breaking Shortcuts via Contrastive Learning. Our second direction is to enhance the agent's SFT/GRPO process via contrastive learning, leveraging negative examples to guide its policy. This involves augmenting the training data with verified conflict samples. For a target scene (e.g., Image_A), we pair it with a mismatched textual prior (e.g., Priors_B) from another scenario whose original ground-truth trajectory (Plan_B) has first been verified via the nuPlan metrics as an unsafe plan for the target scene. For these augmented samples, we will implement a repulsive loss function. This objective penalizes the policy if its predicted trajectory falls within a certain distance margin of the undesirable trajectory, Plan_B. The agent is forced to abandon the "prior equals plan" shortcut and instead learn to genuinely reason, aiming to connect the reasoning and planning.

Furthermore, these two directions are highly complementary. Mitigating modality bias aims to improve the information flowing into the reasoning module, while the contrastive learning strategy aims to enforce that the plan is correctly derived from it. Their combination promises a driving agent whose planning is grounded in its own reasoning.

A.2 Limitation

Our primary fine-tuning experiments employ 7B-parameter models and Omnidrive. Fine-tuning of larger models (e.g., 32B+) remains to be explored. Our diagnostic framework is model-agnostic and can be directly applied to evaluate such models as they become available. Notably, the disconnect already manifests in GPT-4o under zero-shot evaluation, suggesting that model scale alone does not resolve the issue.

A.3 Discussion

The high performance score on benchmarks is not enough. Existing planning metrics mainly evaluate the overlap between predicted and expert trajectories, rather than whether the predicted plan is actually causally grounded in scene understanding. For example, such metrics cannot distinguish genuine reasoning-based planning from shortcut learning. In benchmark settings, this creates a systematic gap from real driving: priors are often strongly correlated with expert trajectories, environmental changes are limited, and performance is judged mainly from final trajectory similarity. Under such conditions, a model can achieve high scores by exploiting priors or memorized trajectory patterns, even without truly understanding the scene.

Our experiments provide direct evidence for this. Even the blind *Plan_NoV* agent, which has no visual input, performs comparably to the full *CoT* agent, while removing priors causes substantial degradation. Moreover, in the Scene Replacement experiment in 4.3, agents still achieve strong planning scores even when the textual priors are paired with completely mismatched visual scenes, where those scores should have dropped sharply if planning were truly scene-grounded. These results show that the apparent 'good performance' is superficial: it reflects weakness of the current evaluation environment rather than genuine safe planning.

The importance of diagnosing the disconnect. Reasoning and planning do need to be aligned for safety in VLM driving models. In autonomous driving carrying human lives, safety is not only

about obtaining a good trajectory in a benchmark; it is life-critical, so being able to verify that the trajectory is produced for the right human-logic reasoning is essential. If reasoning and planning are disconnected, then humans cannot reliably verify whether the vehicle is safe under the logic by which safety should be judged. In that case, an seemingly safe plan may only be a probabilistic coincidence, rather than a safety-guaranteed decision that can be trusted for deployment.

Therefore, the reasoning-planning disconnect is a critical issue precisely because it shows that current "good planning performance" does not come from genuine scene understanding; such performance cannot be trusted as evidence of safe planning in real-world driving. This is why we design the causal probe in this paper: a model should be considered truly strong only if it both achieves high planning scores and passes causal verification showing that its planning is genuinely aligned with scene-grounded reasoning.

Comparison with Other Related Work. Concurrent efforts such as Drive-R1 [22] and Alpamayo-R1 [27] report similar observations to ours and demonstrate improvements in planning performance (e.g., L2 distance and collision rate). These works primarily focus on enhancing model capability, while our work takes a complementary perspective by investigating the underlying reasoning–planning disconnect. In particular, prior studies evaluate improvements through standard planning metrics, but do not explicitly examine whether planning is causally grounded in scene understanding via perturbation or interventional analysis. In contrast, we formally define the reasoning–planning disconnect in VLM driving agents and provide a systematic diagnostic framework to measure it. Our causal probes serve as a validation mechanism to assess whether performance gains are supported by faithful reasoning, offering a foundation that can complement and strengthen future solution-oriented approaches.

A.4 Example data of DriveMind Dataset and Generation Details

This section offers a detailed overview of the DriveMind dataset, covering both its structure and its generation process.

Figure 6 illustrates an example of data from the DriveMind dataset. Our input consists of concatenated surround views and textual inputs (including prior knowledge and instructions). The target output is structured into two parts. First, a plan-aligned CoT is presented within a `<think>` block. This CoT incorporates a comprehensive analysis of the scene, including key static elements (e.g., lanes, traffic lights, road signs), environmental conditions (e.g., weather), and critical dynamic road users (e.g., nearby vehicles, pedestrians, and bicycles), concluding with a macro-level driving decision. Second, following this reasoning, the final trajectory plan is provided within an `<answer>` block. All the reasoning and planning are causally linked through the macro driving decision and the explicit causal logic-enhancing sentence. We highlight in green the information discussed in the main text that may contribute to shortcut learning: ego state, historical data, and navigation information. Our ablation experiments in the main text confirm that VLM driving agents' planning indeed originates from shortcut learning of textual priors, with reasoning emerging as a logical byproduct.

To generate the CoT data, we first use the info parser to extract and make the structured driving info from the nuPlan ground truth log. Then, we use the prompts illustrated in Figure 5 to instruct the GPT-4.1 to generate scenario-aware analysis of the target part, such as weather and detected vehicles. We input the images and corresponding ground truth from the structured info for each part to address the visual-spatial perception limitations inherent in GPT-4.1, and let it generate analyses for each part in sequence. When generating the macro driving decision, all preceding environmental ground truth perceptions, analyses, and the future trajectory are incorporated as inputs, with the given future trajectory ensuring the accuracy of the decision. Notably, to ensure the causality of the reasoning towards driving, we imposed the following constraints on GPT-4.1 when letting it give the macro driving decision: 'While the future trajectory is known, your tone should remain predictive: extrapolating logically from the present and can not report the known future data in your driving decision. Use inferential language, but do NOT use non-predictive language like: according to the future. Your decision must be grounded in present state logic and reflect the causal relationships between the current environment and the driving decision. Remember to remain predictive when extrapolating.' These operations ensure both the accuracy and the causality of our macro driving decision. After all the reasoning generated, we combine them with the ground truth information, plus

Table 6: Comparison of datasets in terms of real-world logs, semantic richness, expert CoT availability, modularity, and scenario coverage.

Dataset	Real-World Logs	Semantic Richness	Expert CoT	Modular Structure	Scenario Coverage
DriveLM	Mixed	Moderate (nuScenes), High (CARLA)	Yes	No	—
DriveVLM	Yes	Moderate (SUP-AD)	Yes	No	40
DriveCoT	No	High (CARLA)	Yes	No	5
OmniDrive	Yes	Moderate (nuScenes)	No	No	—
DriveMind (Ours)	**Yes**	**High (nuPlan)**	**Yes**	**Yes**	**61**

a causal logic-enhancing sentence, to form the CoT data. Finally, a portion of the data will undergo rigorous manual verification to ensure content and logical accuracy.

For the human verification, we randomly sampled 5K samples (10% of the 50K dataset) for human review. Annotators with driving experience assessed each sample on three criteria: (a). accuracy of scene entity descriptions against the visual input, (b). logical correctness of the CoT reasoning chain, and (c). consistency between the reasoning and the ground-truth expert trajectory. Samples failing any criterion were flagged and regenerated. Only 2 out of 5K reviewed samples (0.04%) required regeneration. The extremely low regeneration rate confirms the pipeline produces reliable output at scale, meaning human review can be safely omitted during large-scale expansion.

A key design choice explains the high quality: our pipeline first extracts structured ground truth from nuPlan logs such as 3D object positions, lane topology, traffic light states, etc., and provides all of this to GPT-4.1. With these factual scene elements already established, GPT-4.1's task is narrowly scoped: it only needs to identify the most important causal factor linking these known elements to the given expert trajectory and articulate it as a reasoning chain. This highly constrained generation space leaves minimal room for hallucination or factual error.

A.5 Details of the Scenario Distribution of DriveMind Dataset

We report the exact number of samples associated with each scenario category in the DriveMind dataset, organized by difficulty level:

- Hard: High-dynamic scenarios involving complex interactions (e.g., turning, lane changing).
- General: Standard driving scenarios governed by clear traffic rules (e.g., following lane with lead, stopping at lights).
- Simple: Low-dynamic or static scenarios (e.g., stationary in traffic, low magnitude speed).

The corresponding proportions of scenarios across these difficulty levels are also provided. Our square-root–weighted stratified sampling strategy (see Eq. (1)) yields a well-balanced distribution across difficulty levels, resulting in 33.9% Hard, 39.2% General, and 26.9% Simple scenarios. Detailed statistics are presented in Table 7.

A.6 Comparison with Other Datasets

Table 6 provides a detailed comparison between DriveMind and other VQA datasets in terms of real-world logs, semantic richness, expert CoT availability, modularity, and scenario coverage. DriveMind distinguishes itself from existing datasets through several characteristics:

Real-world fidelity with semantic richness: Unlike simulation-based datasets (e.g., DriveCoT), DriveMind is built upon nuPlan's real-world logs and utilizes its ground-truth annotations with high semantic richness. This not only enriches the environmental semantic information contained within DriveMind, such as stopping line locations and traffic light status, but also ensures the absolute accuracy of the information, both of which significantly ensures the high-quality of DriveMind's VQA.

More diverse scene categories: DriveMind covers 61 distinct driving scenarios, significantly surpassing other datasets such as DriveVLM (40) and DriveCoT (5). Combined with our sampling mechanism, DriveMind incorporates more effective long-tail scenarios and complex scenarios, so that it can provide a more robust foundation for training and evaluating for VLM-based driving agents.

Unique modular structure: DriveMind employs a unique modular design that integrates diverse textual priors, reasoning processes, and planning processes into semantically coherent dialogues. This enables researchers to perform precise information ablation and conduct the causal probe experiments to evaluate a VLM driving model.

High-quality CoT with human verification: DriveMind provides expert CoT annotations generated with assistance from the latest multi-modal LLM GPT-4.1, whereas prior datasets typically rely on older models such as GPT-3.5. In addition, our samples also undergo manual verification, ensuring that the expert CoT annotations are of substantially higher quality.

A.7 GRPO for VLM Driving Agents

This section provides the detailed formulation of the GRPO implementation. We employ three rewards, a location reward, a velocity reward, and a format reward, for our GRPO training. The first two are calculated from the L2 error between each predicted point and its ground truth point, while the latter is consistent with that used in [8]. Note the $K = 3$ rewards as r_1, r_2, r_3 with their weights α_1, α_2, and α_3, the normalized advantage $\mathcal{A}_i$ for an output in one group is calculated as:

$$\mathcal{A}_i = \frac{R_i - \text{mean}(R_1,\ldots,R_G)}{\text{std}(R_1,\ldots,R_G)}, \quad \text{where} \quad R_i = \sum_{k=1}^{K} \alpha_k \cdot r_k. \tag{8}$$

Besides, in equation 3, the $\tilde{w}_i$ is calculated as:

$$\tilde{w}_i = \min\left(\frac{\pi_\theta(o_i|(I_j,T_j))}{\pi_{\theta_{\text{old}}}(o_i|(I_j,T_j))}, \text{clip}\left(\frac{\pi_\theta(o_i|(I_j,T_j))}{\pi_{\theta_{\text{old}}}(o_i|(I_j,T_j))}, 1-\epsilon, 1+\epsilon\right)\right), \tag{9}$$

where the ϵ is a tunable hyperparameter. The KL-divergence penalty item $\mathbb{D}_{\text{KL}}$ is calculated as:

$$\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{\text{ref}}) = \frac{\pi_{\text{ref}}(o_i|(I_j,T_j))}{\pi_\theta(o_i|(I_j,T_j))} - \log\frac{\pi_{\text{ref}}(o_i|(I_j,T_j))}{\pi_\theta(o_i|(I_j,T_j))} - 1. \tag{10}$$

A.8 Tuning Ablation Details

Our tuning ablation experiments are based on our proposed dataset, DriveMind, which contains samples with modular ground truth input and detailed plan-aligned CoT reasoning towards the planning task. We conduct extensive data-driven ablations on different models, verifying the disconnect between reasoning and planning and the shortcut learning leveraging text priors in VLM driving agents. The detailed supplementary information of our ablations is shown in Table 8. All experiments are conducted on 4× NVIDIA A100-80G GPUs. Each SFT run takes approximately 3 days, and each GRPO run takes approximately 5 days. We note that reported times are approximate as experiments are run on a shared cluster.

A.9 Training Details for Agents

In this work, we conduct standard SFT and GRPO training on our agents based on the LLM training and development framework, Scalable Lightweight Infrastructure for Fine-Tuning (SWIFT) [44]. For the SFT, we employ LoRA with rank and alpha both set to 64. Training is conducted on the aligner, ViT, and LLM backbone, using learning rates of 1e-4, 1e-5, and 1e-5, respectively. The batch size is set to 8, with a warm-up ratio of 0.05. Cosine learning rate scheduling is applied to achieve smoother training. For the GRPO training, we employ the same settings, plus the temperature parameter for the policy model that generates samples is set to 0.9 to encourage a degree of exploration, the group size is set to 8, and the warm-up ratio is set to 0.01. For the three rewards, we assign weights of 0.45, 0.45, and 0.1 to the location, velocity, and format rewards for *CoT_grpo*, respectively, whereas for *Base_grpo*, the three rewards are equally weighted.

A.10 nuPlan Challenge Metrics

This section provides detailed definitions for the nuPlan Challenge [26] evaluation metrics used in our experiments.

A.10.1 Open-loop metrics

In each scenario, the agent's planned trajectory is compared against the expert's ground-truth trajectory over 1s, 2s, and 3s horizons. The comparison is based on the following five core metrics, which are first computed at each sampled timestep and then averaged over all timesteps in the scenario.

Table 7: Distribution of scenario types and difficulty-level proportions in the DriveMind.

Difficulty Category	Scenario Name	Sample Count	Total/Proportion
Hard	changing_lane_to_left	56	16935/33.9%
	changing_lane_to_right	51	
	changing_lane_with_trail	24	
	crossed_by_vehicle	41	
	high_lateral_acceleration	648	
	starting_protected_cross_turn	291	
	starting_protected_noncross_turn	434	
	high_magnitude_jerk	33	
	near_barrier_on_driveable	274	
	near_construction_zone_sign	1033	
	near_long_vehicle	1739	
	near_multiple_pedestrians	395	
	near_pedestrian_at_pickup_dropoff	1474	
	near_pedestrian_on_crosswalk	2619	
	near_pedestrian_on_crosswalk_with_ego	267	
	near_trafficcone_on_driveable	1251	
	starting_high_speed_turn	206	
	starting_unprotected_cross_turn	337	
	starting_unprotected_noncross_turn	275	
	traversing_narrow_lane	26	
	behind_long_vehicle	711	
	behind_pedestrian_on_driveable	69	
	behind_pedestrian_on_pickup_dropoff	26	
	behind_bike	205	
	near_multiple_vehicles	1363	
	traversing_pickup_dropoff	3087	
General	near_high_speed_vehicle	1204	19626/39.2%
	waiting_for_pedestrian_to_cross	301	
	accelerating_at_crosswalk	137	
	accelerating_at_stop_sign_no_crosswalk	84	
	accelerating_at_stop_sign	152	
	accelerating_at_traffic_light	100	
	accelerating_at_traffic_light_with_lead	115	
	accelerating_at_traffic_light_without_lead	357	
	following_lane_with_lead	214	
	following_lane_with_slow_lead	828	
	on_all_way_stop_intersection	918	
	on_intersection	1142	
	on_stopline_crosswalk	1014	
	on_stopline_stop_sign	788	
	on_stopline_traffic_light	1342	
	on_traffic_light_intersection	2029	
	starting_low_speed_turn	141	
	starting_straight_stop_sign_intersection_traversal	323	
	starting_straight_traffic_light_intersection_traversal	364	
	stopping_at_crosswalk	219	
	stopping_at_stop_sign_no_crosswalk	100	
	stopping_at_stop_sign_without_lead	166	
	stopping_at_traffic_light_with_lead	75	
	stopping_at_traffic_light_without_lead	122	
	stopping_with_lead	219	
	traversing_crosswalk	1665	
	traversing_intersection	1926	
	traversing_traffic_light_intersection	3581	
Simple	stationary_at_crosswalk	63	13440/26.9%
	stationary_at_traffic_light_with_lead	1206	
	stationary_at_traffic_light_without_lead	1680	
	low_magnitude_speed	1788	
	medium_magnitude_speed	3547	
	following_lane_without_lead	1388	
	high_magnitude_speed	3768	

Table 8: Details of our tuning ablation analysis settings.

Agent	Ablation Details with DriveMind
Base	Base model without any finetuning. Serves as a control group.
CoT	Finetuned on the full DriveMind dataset with ∼50K CoT VQA samples.
Plan	The reasoning part is removed from each sample in DriveMind, retaining only the final planning results.
Plan_NoV	Same as 'Plan', but all visual inputs are also removed.
CoT_NoHis	Ego's history information is removed from the textual input of the samples in DriveMind.
CoT_NoHis_Ego	Both history and the ego state are removed from the textual input of the samples in DriveMind.
CoT_NoPri	All driving priors, history, ego state, and navigation, are removed from the textual input.
CoT_L	Scalability test: Replicates the 'CoT' setup on the LLaVA-1.6.
Plan_L_NoV	Scalability test: Replicates the 'Plan_NoV' setup on the LLaVA-1.6.
Omnidrive	CoT samples from the DriveMind are decomposed into sub-tasks(detection, planning, etc.) without plan-aligned reasoning. Use all the decomposed sub-task data to simultaneously train the Omnidrive agent.
Omnidrive_NoPri	Same as 'Omnidrive', with all the priors removed from text input for each sample.
CoT_grpo	The 'CoT' model is further tuned with GRPO using the $1,000$ grpo VQA samples that have the same input settings as DriveMind.
Base_grpo	The base model is directly tuned with GRPO using the $1,000$ grpo VQA samples.
Base_grpo_NoPri	Same as 'Base_grpo', but all priors in the samples are removed from the text input during GRPO tuning.

Average Displacement Error (ADE). At each timestep, ADE is the mean of the pointwise L2 distances between the planned and expert trajectories over the selected future horizon.

Final Displacement Error (FDE). At each timestep, FDE is the L2 distance between the planned and expert trajectories at the end of the selected future horizon.

Average Heading Error (AHE). At each timestep, AHE is the mean of the absolute differences in heading angle between the planned and expert trajectories over the future horizon.

Final Heading Error (FHE). At each timestep, FHE is the absolute difference in heading angle at the end of the future horizon.

Miss Rate. At each timestep, a "miss" is declared if the maximum pointwise L2 distance between the planned and expert trajectories exceeds a threshold, which is defined as 2.0m for the 1s horizon, 3.2m for 2s, and 6.0m for 3s. The scenario's Miss Rate is the ratio of "missed" timesteps to the total number of timesteps.

Open-Loop Score. The final Open-Loop Score reported in the paper is a composite metric derived from the core metrics via a two-stage process. First, the scenario-level average of each of the five core metrics is compared against a predefined threshold to generate five binary scores ($S_{\text{metric}} \in \{0, 1\}$, where 1 represents a pass). A failure score of 0 is assigned if the scenario's Miss Rate is greater than 30%, if the average ADE or FDE is greater than 8.0 meters, or if the average AHE or FHE is greater than 0.8 radians. These binary scores are then aggregated into a final score ranging from 0 to 100. The S_{MissRate} score acts as a gate; if it is 0, the final score for the scenario is 0. Otherwise, the final score is a weighted average of the other four binary scores, calculated as:

$$\text{Score} = \frac{2 \cdot S_{\text{ADE}} + 2 \cdot S_{\text{FDE}} + S_{\text{AHE}} + S_{\text{FHE}}}{6} \times S_{\text{MissRate}} \times 100 \tag{11}$$

A.10.2 Close-loop metrics

The Closed-Loop Score evaluates the agent's ability to drive safely, adhere to traffic rules, and maintain comfort in an interactive simulation. The score is a composite metric calculated from eight core metrics using a hybrid gating and weighted-average formula. These metrics are divided into two main groups, Gating Metrics and Weighted Metrics.

The first group consists of three gating metrics, which represent critical safety and progress requirements. A failure in any of these results in an immediate scenario score of 0.

No At-Fault Collisions (Collision Ratio). This metric is a three-tiered score. It is assigned a value of 0 for an at-fault collision with a vehicle or Vulnerable Road User (VRU), or for multiple at-fault collisions with objects. It is 0.5 for a single at-fault collision with an object, and 1 otherwise. A

collision is considered "at-fault" if it should have been preventable by the planner, such as colliding with a stopped object or a frontal collision.

Drivable Area Compliance. This is a binary score that becomes 0 if, at any frame, the maximum distance of any corner of the ego's bounding box from the nearest drivable area is more than 0.3m. This tolerance accounts for potential over-approximations of the ego bounding box.

Making Progress. This score is a binary value that is 0 if the agent's progress ratio along the expert route falls below the min progress threshold of 0.2.

The second group comprises five weighted metrics, which are scored between 0 and 1 and contribute to the final performance score via a weighted average.

Driving Direction Compliance. This score is based on the distance traveled against traffic flow over a 1s time horizon; it is 1 if this distance is less than 2m, 0 if it is more than 6m, and 0.5 otherwise.

Time to Collision (TTC). This score is a binary value based on the minimum TTC, which is calculated by projecting bounding boxes forward up to a 3.0s horizon. The score is 0 if the minimum TTC falls below the least TTC safety threshold of 0.95s, and 1 otherwise.

Speed Limit Compliance. This metric evaluates the agent's adherence to posted speed limits from the map data, penalizing both the magnitude and duration of over-speeding. This score is a continuous value between 0 and 1, calculated as: $S_{\text{speed}} = \max\left(0, 1 - \frac{v_{\text{int}}}{v_{\text{thresh}} \cdot T_{\text{scenario}}}\right)$, where v_{int} is the time-integral of the speed violation (i.e., the area under the over-speed vs. time graph), T_{scenario} is the total duration of the scenario, and v_{thresh} is the maximum acceptable over-speeding threshold, set to 2.23 m/s ($\approx$ 5 mph). A score of 1 indicates no violations, and the score trends towards 0 as the severity and duration of over-speeding increase.

Ego Progress along Expert Route. This metric measures how effectively the agent advances along the path taken by the expert driver. At each timestep, the agent's movement is projected onto this path to calculate its per-frame progress. The total progress for both the agent, P_{ego}, and the expert, P_{expert}, is then calculated by integrating these per-frame values over the entire scenario. This score is the ratio of these two values, capped between 0 and 1, and is calculated as:$S_{\text{progress}} = \min\left(1, \frac{\max(P_{\text{ego}}, \tau)}{\max(P_{\text{expert}}, \tau)}\right)$, where τ is a small threshold (0.1m) used to handle minor negative progress values that can arise from data noise and to prevent division by zero in scenarios with no movement. A score of 1 indicates the agent made at least as much progress as the expert.

Comfort. This metric quantifies the smoothness of the planned trajectory by ensuring key variables remain within bounds representative of a comfortable human driving experience. At each timestep of the scenario, a set of quantities are checked against empirically determined thresholds derived from an analysis of expert human driver trajectories. The scenario is deemed uncomfortable, and the score S_{comfort} is set to 0, if *any* of the following conditions are violated at *any* time: longitudinal acceleration is outside the range of [-4.05, 2.40] m/s^2; absolute lateral acceleration exceeds 4.89 m/s^2; absolute yaw rate exceeds 0.95 rad/s; absolute yaw acceleration exceeds 1.93 rad/s^2; absolute longitudinal jerk exceeds 4.13 m/s^3; or the magnitude of the jerk vector exceeds 8.37 m/s^3. If none of these bounds are violated throughout the entire scenario, the score S_{comfort} is 1.

Close-Loop Score. The final Closed-Loop Score is the product of a gating factor (S_{gate}) and a weighted performance score (S_{weighted}). First, the gating factor is the product of the three binary gating metrics, ensuring all critical safety rules are met:

$$S_{\text{gate}} = S_{\text{no_collision}} \times S_{\text{drivable_area}} \times S_{\text{making_progress}} \tag{12}$$

Second, the weighted performance score is the weighted average of the five weighted metrics, with a total weight of 21:

$$S_{\text{weighted}} = \frac{5S_{\text{direction}} + 5S_{\text{TTC}} + 4S_{\text{speed}} + 5S_{\text{progress}} + 2S_{\text{comfort}}}{21} \tag{13}$$

The final Closed-Loop Score is then calculated as follows, where if S_{gate} is 0, the total score is 0:

$$\text{Score} = S_{\text{gate}} \times S_{\text{weighted}} \times 100 \tag{14}$$

A.11 Sequence-level Segment Attention Analysis for Omnidrive

We also conduct the sequence-level attention analysis on the *Omnidrive* model. We primarily select the object detection task (reflecting implicit reasoning processes) and the planning task from the

Table 9: Attention distributions for the *Omnidrive* across preceding sequences during target sequence generation.

Attention Target	Detection Task			Planning Task		
	Shallow Layer	Middle Layer	Final Layer	Shallow Layer	Middle Layer	Final Layer
Image Tokens (%)	**3.46**	**10.02**	**15.92**	1.96	3.22	4.05
Textual Priors (%)	6.76	5.93	6.89	**10.52**	**9.93**	**23.97**
Textual Tokens (%)	96.54	89.98	84.08	98.04	96.78	95.95

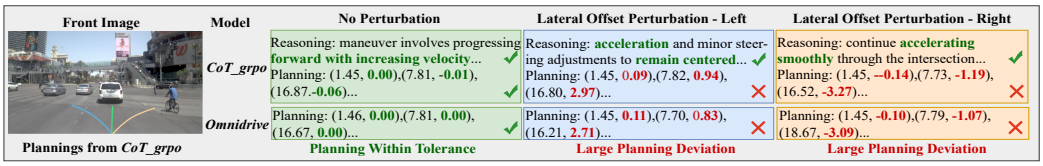


Figure 3: An illustration of our causal probe's diagnostic capabilities in a straight driving scenario. The figure contrasts an agent's original plan with its divergent trajectories when subjected to lateral offset perturbations to the left and right. The lateral trajectories of both models' output undergo a significant change in response to the perturbations.

decomposed CoT subtasks. Since there is no explicit reasoning process, we mainly examine the attention distribution between the sequence of image tokens and the sequence of textual prior tokens in the input across different depth layers during model output.

The results in Table 9 demonstrate that *Omnidrive* progressively focuses more on the image sequence during detection tasks (implicit reasoning). At the final layer, its attention to image tokens reaches 15.92%, while attention to the textual prior sequence fluctuates stably between approximately 6-7%, significantly lower than that for image tokens. This indicates that the textual prior component does not play a decisive role in the detection task that provides the implicit reasoning ability.

During planning tasks, *Omnidrive* consistently allocates significantly high attention weights to textual prior sequences across different layers, ultimately reaching approximately 24%, which is much higher than the attention weights to the image sequence. This demonstrates that *Omnidrive* also heavily relies on textual prior information for planning. The attention weights assigned to the image sequence by *Omnidrive* in the planning task also increase slightly as the layer depth increases. This may stem from the absence of preceding reasoning sequences compared to *CoT_grpo*, resulting in fewer tokens being processed, and consequently increasing attention to images. It may also reflect the model's implicit, superficial and limited processing of image information, as the final attention weight only accounts for 4%.

The results of the sequence-level attention analysis on *Omnidrive* further support our conclusion: current VLM driving agents heavily rely on textual priors during planning, with reasoning processes being almost irrelevant. This also shows that counterfactual reasoning fails to resolve this issue. Furthermore, results also show that modality bias in comprehension persists within *Omnidrive*.

A.12 More Examples of Causal Probe

This section provides more examples of our causal probe applied to diverse driving scenarios, further illustrating the reasoning-planning disconnect and the efficacy of our diagnostic method.

Figure 3 illustrates the application of the causal probe in a simple straight driving scenario. In the baseline case with no perturbation, both the *CoT_grpo* and *Omnidrive* agents correctly produce a plan to continue straight. However, when a minor lateral offset is introduced into the textual priors, both agents exhibit catastrophic planning failures. Their planned trajectories deviate significantly to the left or right, directly following the perturbed history. Crucially, for the *CoT_grpo* agent, the generated reasoning remains correct and consistent across all three cases (e.g., "Reasoning: continue accelerating and minor steer"). Despite the correct reasoning, the planning module produces a wildly divergent and unsafe trajectory. This provides a clear, qualitative example of the planning module ignoring the reasoning module and instead relying on the shortcut provided by the textual priors.

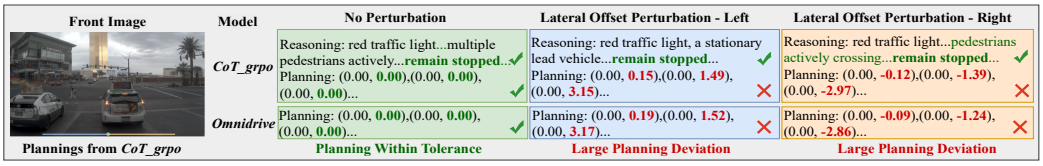


Figure 4: An illustration of our causal probe's diagnostic capabilities in a stopping scenario. The figure contrasts an agent's original plan with its divergent trajectories when subjected to lateral offset perturbations to the left and right. Models output distinct lateral 'translation' trajectories after perturbation, fully demonstrating the disconnect between reasoning and planning, as well as the existence of shortcut learning dependent on text priors.

Figure 4 provides an even more stark example of disconnect in a stationary scenario. In the baseline case, both agents correctly plan to remain stopped at a red light while pedestrians are present. Their reasoning correctly identifies the red light and the need to wait. However, when a lateral offset is applied to the velocity priors, both agents generate an unsafe plan that involves a significant and unnecessary lateral swerve, even while stationary. This case is particularly revealing. The reasoning module for *CoT_grpo* continues to correctly state that the agent should remain stopped due to the red light and pedestrians. Yet, the planning module, completely decoupled from this reasoning, still reacts to the perturbed prior and outputs an erroneous, non-zero motion plan. This demonstrates how deeply ingrained the shortcut learning is, as it can override even a fundamental safety behavior like remaining stopped at a red light.

A.13 Evaluation by Scenario Complexity

To validate the consistency of our findings across different scene complexities, we stratify the 14 distinct evaluation types in the nuPlan challenge test set to the hard, general and simple categories, as stated in Appendix A.5. We measure the performance of the standard agent (*CoT*), the blind agent (*Plan_NoV*), and the prior-deprived agent (*CoT_NoPri*) across these categories. The evaluation results of the 3s planning are summarized in Table 10.

As observed in the table, the "blind" *Plan_NoV* agent achieves parity with (or slightly exceeds) the multimodal *CoT* agent across all complexity levels, including "Hard" scenarios. This confirms that the reasoning-planning disconnect is not limited to simple cases; even in complex turning scenarios, the model ignores visual cues in favor of priors. The performance gap between the standard agent and the prior-removed agent (Δ (*CoT* - *CoT_NoPri*)) serves as a proxy for "reliance on priors." Notably, this gap widens significantly as complexity increases: the Open-Loop score of *CoT_NoPri* drops by 47.16 points in Hard scenarios compared to only 18.50 points in Simple scenarios. These results confirm that the model's reliance on shortcuts is not only consistent but also effectively strengthened under challenging conditions. Rather than engaging in deeper visual reasoning for complex tasks, the agent leans more heavily on textual priors to solve the problem.

A.14 Testing on Extended Training Setups

In this section, we test further training setups to demonstrate that the identified reasoning-planning disconnect is fundamental to existing SFT and RL training, rather than an artifact.

A.14.1 Supervised Fine-tuning: Sequential Training

In our primary evaluation, the *CoT* agent is trained using a joint Supervised Fine-Tuning (SFT) objective, where the model learns to generate both the reasoning steps and the final plan simultaneously given the input (i.e., Input $\rightarrow$ [CoT, Plan]). To investigate whether the observed reasoning-planning disconnect is merely an artifact of this joint training paradigm, we conduct a two-stage sequential training experiment.

In this experiment, we sequentially train two agents with and without text priors, denoted as *CoT_Seq* and *CoT_Seq_NoPri*, using the following two-stage process:

> **Stage 1 (Reasoning Generation):** The model is trained solely to generate the reasoning steps given the input (Input $\rightarrow$ CoT).

Table 10: Open-loop and closed-loop performance comparison across scenario categories.

Category	Scenario	Open-Loop Score ↑			Closed-Loop Score ↑		
		CoT	CoT_NoPri	Plan_NoV	CoT	CoT_NoPri	Plan_NoV
Hard	high_lateral_acceleration	89.69	40.74	90.04	90.31	66.18	90.51
	behind_long_vehicle	98.31	67.25	98.57	99.28	77.74	99.45
	changing_lane	92.33	27.48	92.40	85.96	67.45	86.08
	starting_left_turn	86.98	36.94	87.26	82.93	69.91	83.41
	starting_right_turn	88.64	42.56	88.94	84.63	68.14	85.49
	near_multiple_vehicles	93.42	42.72	93.51	92.79	81.82	92.67
	traversing_pickup_dropoff	89.80	51.35	89.99	84.03	69.27	84.26
Hard Avg		**91.31**	**44.15**	**91.53**	**88.56**	**71.50**	**88.84**
Δ *(CoT - CoT_NoPri)*			**+47.16**			**+17.06**	
General	waiting_for_pedestrian_to_cross	90.14	44.68	90.81	95.27	82.76	95.06
	starting_straight_traffic_light_intersection_traversal	91.84	42.78	92.09	93.57	81.49	93.26
	stopping_with_lead	93.03	67.84	94.87	94.14	81.75	94.35
	following_lane_with_lead	95.16	67.07	96.14	99.98	91.49	99.98
Gen. Avg		**92.54**	**55.59**	**93.48**	**95.74**	**84.37**	**95.66**
Δ *(CoT - CoT_NoPri)*			**+36.95**			**+11.37**	
Simple	low_magnitude_speed	92.71	68.73	92.61	95.26	85.24	95.37
	stationary_in_traffic	99.29	94.47	99.08	99.06	97.74	98.93
	high_magnitude_speed	94.55	67.86	94.56	96.13	87.72	96.78
Simple Avg		**95.52**	**77.02**	**95.42**	**96.82**	**90.23**	**97.03**
Δ *(CoT - CoT_NoPri)*			**+18.50**			**+6.58**	

Table 11: Planning performance (3s) of agents trained with the sequential SFT setup. Removing priors, *CoT_Seq_NoPri* still exhibits a substantial performance degradation.

Agent	Open-Loop Score ↑	Close-Loop Score ↑
CoT_Seq	91.64	91.45
CoT_Seq_NoPri	49.61	74.62

Stage 2 (Planning Generation): A planning module is trained using the input and the model's *own generated* reasoning from Stage 1 to predict the final trajectory ($[\text{Input}, \text{Model's_Own_CoT}] \rightarrow \text{Plan}$).

The results are presented in Table 11. The performance patterns of the sequential agents are similar to the joint SFT setting. The *CoT_Seq* agent achieves high scores, but its planning ability drops drastically when priors are removed (*CoT_Seq_NoPri*), confirming that this two-stage training approach still fails to break the reliance on textual priors. The disconnect is thus fundamental to the existing SFT training or VLM architectural vulnerability, regardless of whether the tasks are learned jointly or sequentially.

A.14.2 RL: GRPO Training with CoT Rewards

In our previous Reinforcement Learning (RL) setup using Group Relative Policy Optimization (GRPO), we primarily employ a 'format' reward. This reward imposes structural discipline on the reasoning process, forcing the model to engage the Chain-of-Thought (CoT) module and adhere to a required structural sequence (e.g., `<think>` → `<answer>`). Though this implicitly encourages the model to analyze the scene from input to output and standardizes the trajectory extraction, here we

Table 12: Performance of agents trained with the visual matching reward under the causal probe ($\delta = 0.1$). Despite high visual matching scores during training, both agents exhibit large deviations.

Agent	MFD ↓	RID ↓
CoT_grpo_visual_match	5.77	1.99
Base_grpo_visual_match	4.42	1.61

design and add a more detailed reward $R_{\text{visual_match}}$, which scores the alignment of the generated CoT with key visual components in the scene:

$$R_{\text{visual_match}} = \frac{\text{count(key_visual_components_in_CoT)}}{\text{count(all_key_visual_components)}}, \tag{15}$$

where 'key visual components' are concretely defined as the critical environmental elements extracted from the ground truth that directly influence driving decisions. Specifically, these include active traffic lights, relevant road signs, and dynamic obstacles (such as pedestrians or vehicles) within a 20m radius. This is to test whether a different reward, such as $R_{\text{visual_match}}$ that directly scores the alignment between the CoT and visual evidence, and a better CoT would mitigate the disconnect.

We train two new agents, *Base_grpo_visual_match* and *CoT_grpo_visual_match*, with this explicitly grounded reward. Both agents successfully learn to articulate visual facts, achieving high average matching scores greater than 0.7. However, we then evaluate these agents under our causal probe ($\delta = 0.1$) to test genuine scene-grounded planning. The results are summarized in Table 12.

As shown in the table, both driving agents still exhibit massive deviations. Even when the model is successfully incentivized to perceive and accurately articulate correct visual facts reasoning (e.g., correctly identifying a red light in text), the planning module still bypasses this reasoning to rely on textual priors. These results show that the reasoning-planning disconnect is a fundamental issue during current RL training, merely a lack of visual attention supervision nor low-quality CoT.

In summary, the reasoning-planning disconnect is a fundamental issue with current VLM driving agents' training. Exploring new SFT paradigms and developing advanced causality rewards to genuinely rebuild the reasoning-planning link remains a critical path for future research. We view those as orthogonal to our current work, and we believe our reasoning-planning decoupling hypothesis and probing tools we have developed are the essential groundwork that future solution-oriented research must build upon.

A.15 Broader Impact Statement

This work identifies a critical safety concern in VLM-based driving agents: their planning may not be causally grounded in reasoning despite appearing interpretable. By exposing this disconnect and providing diagnostic tools, we aim to promote more robust development and evaluation practices for autonomous driving systems before real-world deployment. We think there are no negative societal impacts of this diagnostic research.

A.16 The Use of Large Language Models

In preparing this manuscript, we utilized an LLM (i.e., GPT-5) as a tool to enhance the clarity and precision of our writing. The core ideas, arguments, and scientific contributions are conceived and authored by the human writers. The LLM was employed to refine sentence structure, improve readability, and ensure grammatical correctness on initial human-written drafts. The final version of the manuscript reflects the authors' thorough review and final judgment on all text.

Prompt for Traffic Light Reasoning:

<If the light stays the same, what can ego-vehicle do? If it changes, what can ego-vehicle do? Take into account the surrounding environment and the ego-vehicle's current state.>

Prompt for Traffic Sign Reasoning:

<Identify any traffic signs visible in the front views, describe their meaning, and explain whether they affect the current movement of the ego-vehicle. If no signs are detected, state that no traffic signs are visible ahead.>

Prompt for Weather Reasoning:

<Give a short description of whether in the current driving scenario, then briefly analyze if the weather poses a threat to the ego-vehicle's current driving, and if it poses a threat, what should the ego-vehicle do.>

Prompt for Detected Vehicle Reasoning:

<Briefly analyze the vehicle's intent and whether the vehicle may threaten the driving of the ego-vehicle, and if it poses a threat, analyze what kind of threat it would be. Your analysis should refer to the 'Addition info' but should not include the 'Addition info'. 'Addition info: The position of the object in the past 1 second is ({x_past:.2f},{y_past:.2f}).>

Prompt for Detected Pedestrian Reasoning:

<Briefly analyze the motion of the pedestrian, then analyze whether the pedestrian may threaten the driving of the ego-vehicle, and if it poses a threat, analyze what kind of threat it would be. Your analysis should refer to the 'Addition info' but should not include the 'Addition info'. Addition info: The position of the object in the past 1 second is ({x_past:.2f},{y_past:.2f}>

Prompt for Detected Bicycle Reasoning:

<Briefly analyze the bicycle's intent and whether it may threaten the driving of the ego-vehicle, and if it poses a threat, analyze what kind of threat it would be. Your analysis should refer to the 'Addition info' but should not include the 'Addition info'. Additional info: The position of the object in the past 1 second is ({x_past:.2f},{y_past:.2f}).>

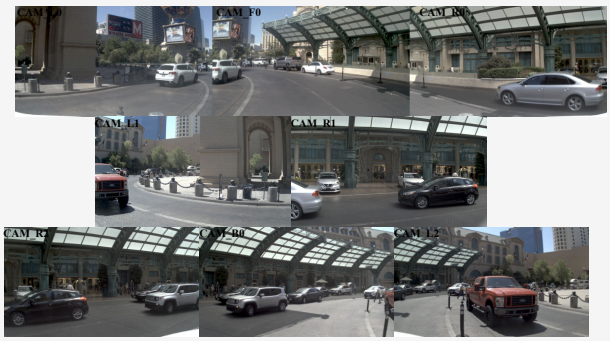
Prompt for Driving Decision Reasoning:

<Give a brief driving decision based on all the current information and analysis, and common sense, traffic principles in driving. It should clearly guide the ego-vehicle's action over the next 3 seconds. You should specify the anticipated maneuver, provide the anticipated changes in ego-vehicle's position (x, y) and velocity (vx, vy) according to the maneuver in the near future 3 seconds and the reasons.

While the future trajectory is known, your tone should remain predictive: extrapolating logically from the present and can not report the known future data in your driving decision. Use inferential language, but do NOT use non-predictive language like: according to the future. Your decision must be grounded in present state logic and reflect the causal relationships between the current environment and the driving decision. Remember to remain predictive when extrapolating.>

Figure 5: Prompt for GPT-4.1 to generate analysis for different parts. All together combine to form the plan-aligned CoT in DriveMind.

Input Multi-view Images:



Input Text Prompt:

Here are three stitched images representing the current driving scene, captured from eight cameras mounted on the ego-vehicle that you are driving:
-Image 1 (left to right): Front-left (CAM_L0), Front (CAM_F0), Front-right (CAM_R0)
-Image 2 (left to right): Left (CAM_L1), Right (CAM_R1)
-Image 3 (left to right): Rear-right (CAM_R2), Rear (CAM_B0), Rear-left (CAM_L2)
The ego-vehicle your are driving has the dimension (length, width and height): (5.18,2.30,1.78).
Current ego-vehicle state: location (0.0, 0.0), velocity (3.71,-0.10), total velocity magnitude 3.72.
Lane speed limit: 3.24m/s.
Navigation: Center point of the next target road block: (29.58, 10.62).
Past 3 seconds ego-vehicle states:
- at past 3.0 s: location (-10.39, 1.22), velocity (3.21, -1.27), total velocity magnitude 3.45...
- at past 2.0 s: location (-7.14, 0.45), velocity (3.37, -0.85), total velocity magnitude 3.47...
- at past 1.0 s: location (-3.66, 0.06), velocity (3.75, -0.45), total velocity magnitude 3.78...
Your task is to give the trajectory and speed of the ego-vehicle for the next 3 seconds at 0.2 second prediction intervals based on the current camera data, known information about the ego-vehicle, and navigation information. Please analyze the driving environment first, you should consider aspects as follows step by step: Lane Analysis, Traffic Light Analysis, Traffic Sign detection ahead the ego-vehicle, Weather Analysis, Detected critical Vehicles, Detected critical Pedestrians, Detected critical bicycles, Detected other objects. Then, based on all the information, combined with Common Sense in driving, Traffic Principles and reasonable assumptions about the future, give a macro driving decision that decides how the ego-vehicle should act over the next 3 seconds. Finally, give the predicted ego-vehicle states—location (x, y) and velocity (vx, vy)—over the next 3.0 seconds. All of your analysis including the macro driving decision should be enclosed in a pair of <think></think>, and your final predictions should be enclosed in a pair of <answer></answer>.

Ground Truth plan-aligned CoT Answer:

<think>
Now I'm driving in the lane.
Lane Analysis:
2 lanes detected from ego-vehicle's cameras. Taking the heading direction of the ego-vehicle and the location of the ego-vehicle as references, lanes in order from left to right are:
- Lane on the left side, same direction as the ego-vehicle's heading. Lane width: 4.35.
- Lane the ego-vihicle in, same direction as the ego-vehicle's heading. Lane width: 4.95.
Traffic Light Analysis:
No traffic lights detected in the current scenario.
Traffic Sign detection ahead the ego-vehicle:
No traffic signs are visible ahead.
Weather Analysis:
It is daytime with clear weather, as evidenced by bright sunlight, clear shadows, and good visibility in the images. The weather does not pose any threat to the ego-vehicle's driving.
Detected critical Vehicles:
-a silver sedan with a black roof and a silver logo on the front hood, can be seen from ['CAM_R0', 'CAM_R1'], now at location (2.62,-6.57), at the front-right of the ego vehicle, dimension (4.92,1.87,1.53). The vehicle seems moving slowly forward and very slightly to the right. It is likely waiting, parked, or maneuvering in a drop-off area and does not currently threaten the ego-vehicle, provided the ego-vehicle remains in its lane.
-a black hatchback with a logo on the front hood, can be seen from ['CAM_R1', 'CAM_R2'], now at location (-5.30,-6.74), at the rear-right of the ego vehicle, dimension (4.37,1.83,1.51). The vehicle seems slowly moving forward and slightly to the right. It seems to be maneuvering through a congested drop-off area, and is not actively threatening the ego-vehicle as it stays well behind and to the right.
-an orange pickup truck with large tires, can be seen from ['CAM_B0', 'CAM_L1', 'CAM_L2'], now at location (-8.34,5.92), at the rear-left of the ego vehicle, dimension (6.45,2.24,2.20). The vehicle is moving very slowly forward and slightly left. Its position suggest it is unlikely to enter the ego-vehicle's path, so it does not pose a current threat.
Detected critical Pedestrians:
-a pedestrian, can be seen from ['CAM_B0', 'CAM_L2'], now at location (-16.40,8.26), at the rear-left of the ego vehicle, dimension (1.04,0.87,1.92). The pedestrian is moving forward and significantly toward the left, but is located far behind the ego-vehicle and in a separate area, presenting no threat.
Detected critical bicycles:
- No bicycle detected.
Detected other objects:
- an object, at location (1.58,1.96), at the front-left of the ego vehicle, dimension (0.44,0.45,1.22).
- an object, at location (-5.95,2.72), at the rear-left of the ego vehicle, dimension (0.43,0.44,1.23).
Macro driving decision:
The ego-vehicle is currently traveling at a speed slightly above the lane speed limit, progressing in a gentle leftward curve as seen in the current road geometry. There are no obstacles or active threats detected in the immediate path, with all identified vehicles and pedestrians remaining clear of the driving lane. The environment is controlled and slow-paced, typical of a hotel drop-off or pickup loop with some stationary or slow vehicles and scattered pedestrians. Over the next 3 seconds, the ego-vehicle should anticipate a continued gentle left curve to remain centered in the lane while gradually adjusting speed closer to the posted limit to ensure both compliance and safety. The vehicle is expected to shift incrementally leftward (increasing positive y), as demonstrated by recent velocity changes, while longitudinal velocity (vx) gently decreases as it adapts for the curve and aligns with lane limits. This maneuver promotes lane-keeping, smoothness, and readiness for any evolving activity in the area ahead. The vehicle should stay alert for possible merging or entering traffic near the curbside but does not need to take evasive action given the current clear path.
</think>

<answer>
Based on current environmental perception with comprehensive analysis, driving decision, and reasonable assumptions about the future, the predicted ego-vehicle states—location (x, y) and velocity (vx, vy)—over the next 3.0 seconds are as follows:\n- at future 0.2 s: location (0.72, 0.04), velocity (3.69, 0.01), total velocity magnitude 3.69.\n- at future 0.4 s: location (1.44, 0.09), velocity (3.65, 0.07), total velocity magnitude 3.65.\n- at future 0.6 s: location (2.15, 0.15), velocity (3.62, 0.13), total velocity magnitude 3.63.\n- at future 0.8 s: location (2.84, 0.22), velocity (3.58, 0.21), total velocity magnitude 3.59.\n- at future 1.0 s: location (3.53, 0.32), velocity (3.54, 0.28), total velocity magnitude 3.55.\n- at future 1.2 s: location (4.21, 0.43), velocity (3.50, 0.35), total velocity magnitude 3.52.\n- at future 1.4 s: location (4.87, 0.55), velocity (3.45, 0.42), total velocity magnitude 3.48.\n- at future 1.6 s: location (5.52, 0.68), velocity (3.40, 0.51), total velocity magnitude 3.43.\n- at future 1.8 s: location (6.16, 0.83), velocity (3.35, 0.56), total velocity magnitude 3.39.\n- at future 2.0 s: location (6.79, 0.98), velocity (3.30, 0.63), total velocity magnitude 3.36.\n- at future 2.2 s: location (7.42, 1.15), velocity (3.26, 0.67), total velocity magnitude 3.33.\n- at future 2.4 s: location (8.03, 1.32), velocity (3.21, 0.73), total velocity magnitude 3.29.\n- at future 2.6 s: location (8.64, 1.50), velocity (3.16, 0.79), total velocity magnitude 3.26.\n- at future 2.8 s: location (9.24, 1.69), velocity (3.12, 0.82), total velocity magnitude 3.23.\n- at future 3.0 s: location (9.82, 1.89), velocity (3.09, 0.87), total velocity magnitude 3.21.
</answer>

Figure 6: An example of our plan-aligned CoT driving VQA data from DriveMind, with the three priors in the text input marked as green.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state our contributions and research scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Appendix A.2 and referenced in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper does not include theoretical proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The methodology is described in detail in Section 3. Training details, hyperparameters, and evaluation protocols are provided in Appendix A.9 and A.10.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and data will be progressively made available in the public repository linked on the first page of the paper.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training details and hyperparameters are in Appendix A.9; evaluation protocol is in Appendix A.10.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Our evaluation is conducted on the nuPlan fixed test set with deterministic inference.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Compute resources are reported in Appendix A.8 and A.9.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: A broader impact statement is provided in Appendix A.15.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The datasets and models used in this work pose no risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All existing assets (nuPlan, Qwen2.5-VL, LLaVA, etc.) are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The DriveMind dataset is documented in detail in Section 3, with generation pipeline, data examples, distribution statistics, and comparisons provided in the Appendix A.4, A.5, A.6.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve human subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: GPT-4.1 is used for generating plan-aligned CoT annotations in our data pipeline, as described in Section 3.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.