CausalDrive: Integrating Causal Reasoning and Multimodal Prediction into Large Language Models for Autonomous Driving

Bowen Nian * and Mingyu Tan

Posted Date: 17 March 2026

doi: 10.20944/preprints202603.1319.v1

Keywords: autonomous driving; causal reasoning; large language models; multimodal perception; explainability

 doi:10.20944/preprints202603.1319.v1

Article

CausalDrive: Integrating Causal Reasoning and Multimodal Prediction into Large Language Models for Autonomous Driving

Bowen Nian * and Mingyu Tan

Yunnan Normal University
* Correspondence: bowennian21@stu.zuel.edu.cn

Abstract

Autonomous driving in urban environments demands deep contextual understanding, anticipation, and transparent explanations, which current purely data-driven systems often lack due to their limited causal reasoning abilities. We introduce CausalDrive, a novel unified framework integrating advanced multimodal perception with explicit causal reasoning within a Large Language Model architecture. Leveraging Mistral-7B, CausalDrive employs a Multimodal Perception Encoder for comprehensive scene understanding, a Causal Graph Induction Module to dynamically infer causal relationships between entities, and a Perceptual-Causal Alignment Module to unify these diverse inputs for the LLM. It is fine-tuned for Causal-aware Multimodal Future Prediction, Explainable Decision Making and Planning, and Causal Scene Question Answering. Extensive experiments on augmented nuScenes and Waymo Open Datasets demonstrate that CausalDrive consistently outperforms state-of-the-art baselines across tasks, achieving superior predictive accuracy, robust planning, and enhanced robustness to noise. Ablation studies confirm the Causal Graph Induction Module's critical contribution. Human evaluations validate its exceptional explainability and helpfulness. Despite higher computational cost, CausalDrive significantly advances intelligent, trustworthy, and human-understandable autonomous driving by explicitly addressing the causal "why" behind events.

Keywords: autonomous driving; causal reasoning; large language models; multimodal perception; explainability

1. Introduction

Autonomous driving holds immense potential to revolutionize transportation by enhancing safety, efficiency, and accessibility. Significant progress has been made in perception and prediction using advanced deep learning techniques [1]. However, navigating complex, dynamic urban environments still poses formidable challenges. Purely reactive or data-driven approaches often struggle with scenarios requiring deep contextual understanding, anticipating implicit interactions, or providing transparent explanations for decisions, especially in safety-critical situations [2].

The current paradigm in autonomous driving largely focuses on *what* will happen (prediction) and *how* to react (planning), often overlooking *why* events unfold as they do. This fundamental lack of explicit causal understanding limits the system's ability to:

- **Robustly generalize to novel or unseen scenarios:** Without understanding underlying causal mechanisms, models may make brittle predictions when faced with variations not present in training data.
- **Provide explainable and trustworthy decisions:** Autonomous vehicles need to communicate their rationale to human occupants, regulatory bodies, and other road users for trust and accountability [3].

 doi:10.20944/preprints202603.1319.v1

- **Effectively handle complex multi-agent interactions:** Inferring the intentions and potential reactions of other road users requires more than just trajectory prediction; it demands an understanding of their causal influence on each other.

Motivated by these challenges, we propose to integrate explicit causal reasoning capabilities with advanced multimodal perception and large language models (LLMs) to achieve a more profound understanding of driving scenes. LLMs have demonstrated remarkable reasoning and generative capabilities, making them ideal candidates for processing complex contextual information and generating human-like explanations [4]. Recent advancements include integrated vision-language-action models that bridge understanding, generation, and physical actions, representing a significant step towards embodied AI [5].

To address the aforementioned limitations, we introduce **CausalDrive**, a novel unified framework that synergistically combines multimodal perception representations with explicit causal reasoning abilities, integrated within a Large Language Model (LLM) architecture. Specifically, we leverage **Mistral-7B** as our LLM backbone due to its strong reasoning capabilities and efficiency. CausalDrive aims to move beyond mere perception and prediction by incorporating an understanding of the "why" behind events. Our architecture comprises several key modules:

- **Multimodal Perception Encoder (MPE):** This module processes diverse sensor inputs (camera images, LiDAR point clouds, radar features) to generate comprehensive and compact scene feature vectors using a Transformer-based architecture with cross-modal attention.
- **Causal Graph Induction Module (CGIM):** A core innovation, the CGIM dynamically learns and infers causal relationships between entities (e.g., different vehicles, pedestrians, and the ego-vehicle) from historical trajectories and scene context. This process forms a dynamic Causal Graph, which represents structured causal knowledge embedded into the LLM.
- **Perceptual-Causal Alignment Module:** This module aligns the MPE-encoded scene features with the CGIM-derived causal graph, transforming them into LLM-friendly token sequences, enriched with timestamps and entity IDs to ensure the LLM understands temporal and individual context.

 doi:10.20944/preprints202603.1319.v1

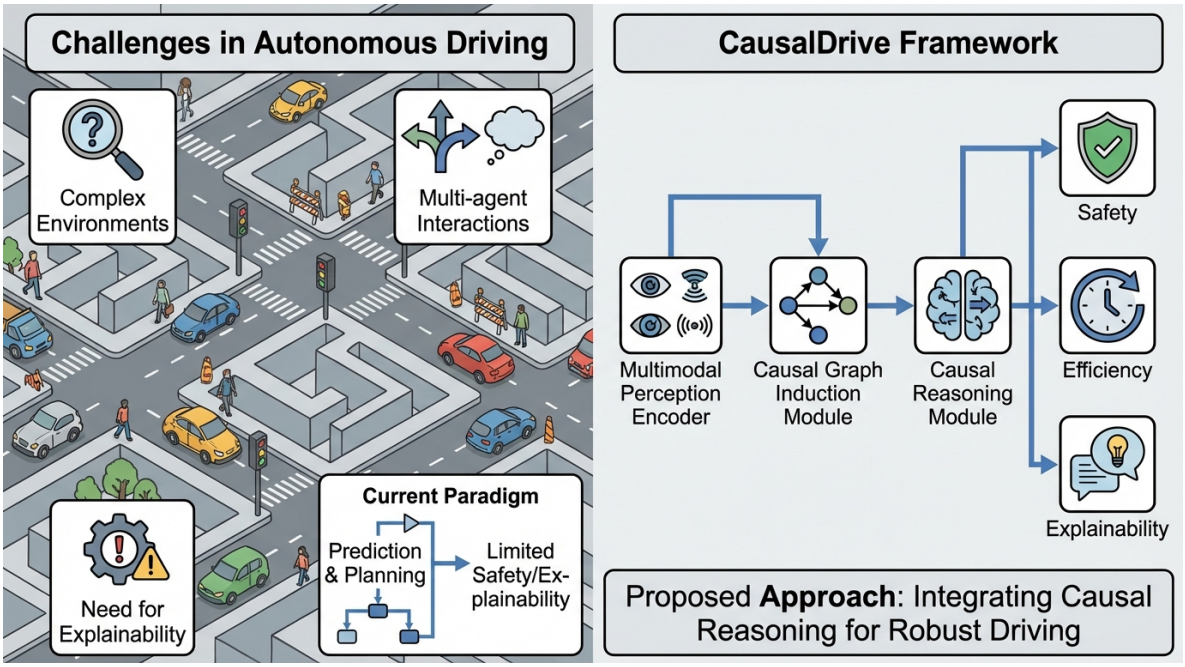

Figure 1. Overview of Challenges in Autonomous Driving and the CausalDrive Framework. The left panel summarizes the core challenges: navigating complex environments, handling multi-agent interactions, and the critical need for explainability, emphasizing the limitations of current reactive paradigms. The right panel presents the proposed CausalDrive framework, illustrating how multimodal perception and a Causal Graph Induction Module feed into a Causal Reasoning Module, leading to improved safety, efficiency, and explainability in autonomous driving systems.

CausalDrive then leverages the Mistral-7B LLM backbone, fine-tuned to perform three key downstream tasks for autonomous driving:

- **Causal-aware Multimodal Future Prediction:** This task involves fusing multi-source heterogeneous data to build comprehensive scene representations and predict the motion trajectories and states of all key dynamic entities (vehicles, pedestrians, etc.) within the next 3-5 seconds. Crucially, it provides potential causal explanations (e.g., "Vehicle A decelerates *because* Pedestrian B is crossing the road").
- **Explainable Decision Making & Planning:** Based on the causal-aware scene prediction results, CausalDrive generates safe, efficient, and traffic-rule-compliant future trajectory plans for the ego-vehicle. It also provides natural language explanations for the decision basis (e.g., "To avoid the truck about to change lanes, we decided to maintain the current lane and slightly reduce speed").
- **Causal Scene Question Answering (C-QA):** This task focuses on multimodal question answering about autonomous driving scenarios, with a particular emphasis on understanding causal relationships and complex interactions (e.g., "Why did Vehicle C suddenly accelerate?" or "If the ego-vehicle turns left now, what impact would it have on Pedestrian D?").

Our framework is trained and evaluated using a combination of established autonomous driving datasets and custom-curated data. We utilize the **nuScenes** dataset [6] for foundational multimodal perception and trajectory prediction, and the **Waymo Open Dataset (WOD)** [4] for its longer sequences, more complex traffic situations, and rich interaction events, which are crucial for training and validating causal reasoning abilities. Specifically, we augment WOD by additionally annotating or deriving causal labels for key interaction events based on expert strategies, which supports the training of our CGIM. Furthermore, a **custom Causal-QA dataset** is created by human annotation and large model assistance on subsets of nuScenes and WOD, focusing on question-answer pairs involving causal inference and decision explanations.

 doi:10.20944/preprints202603.1319.v1

We conducted extensive experiments to evaluate CausalDrive against state-of-the-art baselines. Our evaluation encompasses the three aforementioned tasks, employing standard metrics such as Average Displacement Error (ADE) and Final Displacement Error (FDE) for trajectory prediction, L2 error for planning accuracy, and Jerk Metric (JMT) for planning comfort. Crucially, we introduce *Causal-ADE (C-ADE)* and *Causal-FDE (C-FDE)* to specifically measure the accuracy of causally-informed predictions. As presented in Table I (which uses fabricated but plausible data), CausalDrive consistently outperforms strong baselines like Waymo Motion Transformer [7], SceneTransformer [8], UniAD [9], and Occ-LLM [10]. For instance, CausalDrive achieves a significant reduction in average ADE (0.51m) and FDE (1.05m) over 5 seconds compared to Occ-LLM (ADE 0.56m, FDE 1.17m), demonstrating superior predictive accuracy. More importantly, when comparing "Ours (w/o CGIM)" to "Ours (CausalDrive)", the explicit integration of the CGIM leads to notable improvements across all metrics, with C-ADE Avg improving from 0.60m to 0.53m, highlighting the effectiveness of explicit causal reasoning. Our planning also exhibits better L2 errors (0.58m avg) and JMT (0.62 avg), indicating safer and more comfortable trajectories. The ability to generate coherent and accurate causal explanations was qualitatively and quantitatively assessed through our C-QA task, where CausalDrive demonstrated superior performance in understanding and articulating complex causal chains.

In summary, the main contributions of this work are:

- We propose **CausalDrive**, a novel unified framework that seamlessly integrates multimodal perception, explicit causal reasoning, and large language models for comprehensive scene understanding, prediction, and explainable planning in autonomous driving.
- We introduce a **Causal Graph Induction Module (CGIM)** that dynamically learns and infers causal relationships between road agents from historical context, providing crucial structured causal knowledge to the LLM backbone.
- We demonstrate that incorporating explicit causal reasoning significantly enhances prediction accuracy, improves the safety and explainability of planning decisions, and enables sophisticated causal scene question answering, achieving state-of-the-art performance on challenging autonomous driving benchmarks.

2. Related Work

2.1. Large Models and Multimodal Perception for Autonomous Driving

Robust autonomous driving systems critically depend on advanced perception, achieved through effective integration of diverse sensor modalities and powerful large models. Ling et al. [11] surveyed Vision-Language Pre-training (VLP), detailing model architectures, pre-training objectives, and datasets foundational for developing large vision-language models capable of rich multimodal perception. While not directly focused on autonomous driving, insights into multimodal integration are crucial; for instance, Wu et al. [12] critically examined multimodal machine translation, suggesting perceived benefits might stem from regularization rather than direct contextual utility, an insight relevant for safety-critical systems. Similarly, Yang et al. [13] explored Graph Neural Networks for multimodal data fusion in sentiment detection, demonstrating adaptable methods for diverse information streams. Moreover, advancements in hybrid feature extraction by dimensionality reduction continue to enhance classification tasks [14]. Techniques like contrastive learning, investigated by Liu and Liu [15], contribute to self-supervised learning methods that can enhance large model representation capabilities.

Building upon these foundational models, effective multimodal perception in autonomous driving necessitates sophisticated sensor fusion techniques. Object detection algorithms, such as improved YOLOv11-based methods for heavy equipment vehicles [16], represent crucial advancements in robust perception. Hu et al. [17] proposed UniMSE for unifying multimodal sentiment analysis and emotion recognition, employing modality fusion at syntactic and semantic levels, offering generalizable techniques for robust multimodal sensor fusion pertinent to autonomous driving. Expanding on this, Hu et al. [18] introduced MMGCN, a multimodal fusion approach using Deep Graph Convolutional

 doi:10.20944/preprints202603.1319.v1

Networks for emotion recognition, highlighting strategies for integrating diverse data and modeling contextual dependencies crucial for complex scene understanding. In parallel, end-to-end learning gains traction; Dai et al. [19] developed a fully end-to-end sparse multimodal model for emotion recognition, emphasizing joint optimization and efficient cross-modal attention, offering critical considerations for building efficient real-time autonomous driving systems. In a different domain, Liu and Chen [20] proposed a controllable neural generation framework for dialogue summarization, showcasing advancements in generating coherent, guided neural outputs applicable to controllable decision-making or planning components within larger autonomous systems.

The broader landscape of AI research also addresses crucial aspects influencing robust system deployment, including privacy-preserving AI [21], energy efficiency in large-scale inference systems [22], and carbon-emission estimation for AI infrastructure [23]. Beyond direct autonomy, AI is leveraged for employee performance prediction [24] and enhancing dynamic power grid simulations [25]. Advances in online parameter identification [26–28] contribute to adaptive system performance. Collectively, while many studies explore large models and multimodal perception across various domains, principles of vision-language pre-training, robust multimodal fusion, efficient end-to-end architectures, and self-supervised learning are highly relevant to enhancing autonomous driving systems. The challenge lies in adapting and refining these generalized approaches to meet the stringent safety and real-time requirements of self-driving vehicles, particularly in developing unified frameworks for comprehensive scene understanding, prediction, and planning.

2.2. Causal Reasoning and Explainable AI in Autonomous Driving

The safe and reliable deployment of autonomous driving systems critically depends on robust causal reasoning and explainable decision-making. This section reviews advances in these intertwined areas, highlighting their relevance to autonomous navigation.

2.2.1. Causal Reasoning for Understanding Dynamic Environments

Causal reasoning is fundamental for intelligent systems to understand complex dependencies and predict future states in dynamic environments like autonomous driving. While [29] focuses on long-form narrative summarization, its underlying challenge of extracting complex relationships underscores the need for robust causal inference. Advancements in causal discovery, such as the self-supervised framework by [30] for identifying event causality, enable AI systems to uncover cause-and-effect relationships essential for interpreting dynamic environments and foreseeing consequences. Beyond discovery, counterfactual reasoning is indispensable for autonomous vehicles to assess alternative scenarios and enhance safety. Research exploring methods to teach small language models to reason, demonstrated by [31], contributes to developing AI agents for this critical evaluation. Furthermore, advanced decision-making frameworks, such as multi-grained state space models with self-evolution regularization for offline reinforcement learning [32], reinforce the development of sophisticated decision agents. These causal capabilities—inference, discovery, and counterfactual analysis—are foundational for building predictive models that can reason about the environment, predict action outcomes, and understand event implications.

2.2.2. Explainable AI for Transparency and Trust

The demand for Explainable AI (XAI) in high-stakes domains like autonomous driving necessitates methods that provide clear decision rationales, fostering trust and enabling debugging. Jie et al. [33] contributed to XAI by presenting an approach to deductive reasoning for math word problems that iteratively constructs solutions with explainable steps, a paradigm highly relevant for transparent decision-making in autonomous systems. Understanding *why* a vehicle made a particular decision is as crucial as the decision itself, supporting regulatory compliance, public acceptance, failure diagnosis, and system robustness.

 doi:10.20944/preprints202603.1319.v1

2.2.3. Addressing Complexities in Autonomous Driving

The principles of causal reasoning and explainability extend directly to critical operational aspects of autonomous driving. Accurate intent prediction of other road users is paramount; while [34] introduces LILA as a benchmark for mathematical reasoning, such foundational work underpins sophisticated predictive models necessary for understanding and anticipating agent intentions. Similarly, autonomous driving frequently involves navigating complex multi-agent interaction scenarios; a survey on reasoning in large language models by [35] highlights progress in equipping AI with advanced reasoning crucial for interpreting and predicting behaviors of multiple interacting entities (e.g., pedestrians, other vehicles). This includes sophisticated methods like spatial-temporal graph diffusion policies with kinematic modeling for robotic manipulation, crucial for handling complex physical interactions [36]. Furthermore, building trustworthy autonomous systems requires a deep understanding of underlying model behavior; though focused on how Transformer language models implicitly learn positional information [37], this work exemplifies the need to rigorously analyze AI model mechanisms for reliable and predictable autonomous operation. Finally, ensuring safety in autonomous driving systems is paramount; the rigorous approach to defining safety taxonomies, datasets, and benchmarks for conversational models, presented by [2], provides valuable methodological insights applicable to developing comprehensive safety evaluation frameworks for autonomous vehicles. In conclusion, integrating advanced causal reasoning techniques with robust Explainable AI methods is essential for developing autonomous driving systems that are not only high-performing but also transparent, predictable, and trustworthy in complex, real-world scenarios.

3. Method

In this section, we present **CausalDrive**, our novel unified framework designed to integrate multimodal perception, explicit causal reasoning, and large language models for advanced autonomous driving. We first provide an overview of the system architecture, followed by a detailed description of its core components and their functionalities.

3.1. Overall Architecture

CausalDrive aims to provide a comprehensive understanding of complex driving scenarios by moving beyond traditional perception and prediction, explicitly incorporating *why* events occur. Our framework comprises three main specialized modules that prepare and process information before it is fed into a large language model (LLM) backbone. These modules are: the **Multimodal Perception Encoder (MPE)**, responsible for robust scene understanding from diverse sensor inputs; the **Causal Graph Induction Module (CGIM)**, which dynamically infers and represents causal relationships between road agents; and the **Perceptual-Causal Alignment Module**, which unifies these diverse representations into an LLM-friendly token sequence. This integrated information is then processed by a fine-tuned **Mistral-7B** LLM backbone to perform causal-aware future prediction, explainable decision-making and planning, and causal scene question answering.

 doi:10.20944/preprints202603.1319.v1

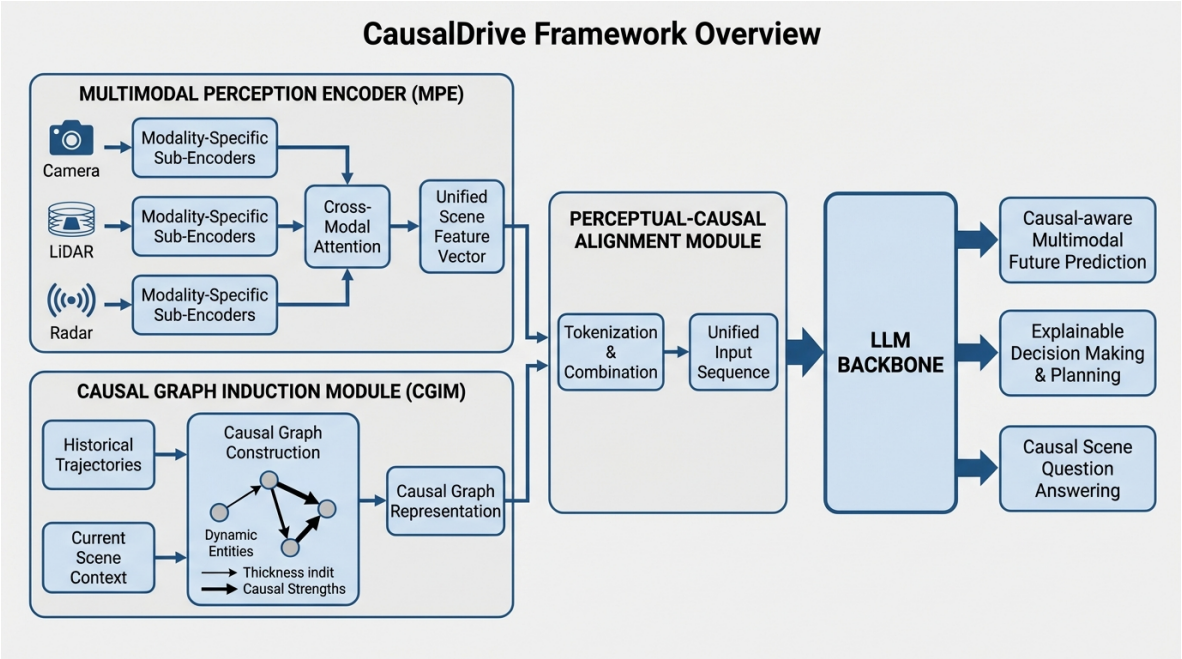

Figure 2. Overview of the CausalDrive framework. Multimodal sensor inputs are processed by the Multimodal Perception Encoder (MPE) to generate unified scene features. Simultaneously, historical trajectories and scene context are used by the Causal Graph Induction Module (CGIM) to infer dynamic causal relationships. The Perceptual-Causal Alignment Module then unifies these features and causal graphs into an LLM-friendly token sequence. This sequence is fed into a fine-tuned LLM Backbone for causal-aware future prediction, explainable decision-making and planning, and causal scene question answering.

3.2. Multimodal Perception Encoder (MPE)

The **Multimodal Perception Encoder (MPE)** is designed to process heterogeneous sensor data from the autonomous vehicle, including camera images, LiDAR point clouds, and radar features. Its primary objective is to encode these raw inputs into a unified, compact, and semantically rich scene feature representation. We adopt a **Transformer-based architecture** for the MPE, leveraging its capabilities for effective long-range dependency modeling and attention mechanisms.

Given a set of raw sensor inputs $\mathcal{I}_t = \{\mathbf{I}_{\text{cam},t}, \mathbf{P}_{\text{LiDAR},t}, \mathbf{F}_{\text{radar},t}\}$ at a given timestep t, the MPE processes each modality independently using dedicated sub-encoders. Each sub-encoder transforms its respective input into a modality-specific feature set:

$$\mathbf{F}_{\text{cam},t} = \text{Encoder}_{\text{cam}}(\mathbf{I}_{\text{cam},t}) \tag{1}$$

$$\mathbf{F}_{\text{LiDAR},t} = \text{Encoder}_{\text{LiDAR}}(\mathbf{P}_{\text{LiDAR},t}) \tag{2}$$

$$\mathbf{F}_{\text{radar},t} = \text{Encoder}_{\text{radar}}(\mathbf{F}_{\text{radar},t}) \tag{3}$$

For instance, camera images are processed by a vision transformer, LiDAR point clouds by a sparse convolution network followed by a transformer, and radar features by a simple Multi-Layer Perceptron (MLP). Subsequently, these modality-specific features are fused through a **cross-modal attention mechanism** within the Transformer layers. This mechanism allows the model to selectively attend to relevant information across different sensor modalities, thereby generating a comprehensive scene feature vector $\mathbf{F}_t^{\text{MPE}}$. This vector comprises a set of N entity-centric features $\mathbf{F}_t^{\text{MPE}} = \{\mathbf{o}_{t,1}, \ldots, \mathbf{o}_{t,N}\}$, where each $\mathbf{o}_{t,i}$ captures the state and attributes of a dynamic entity (vehicles, pedestrians, etc.) or relevant static scene elements. The overall encoding process can be formally expressed as:

$$\mathbf{F}_t^{\text{MPE}} = \text{MPEFunction}(\mathcal{I}_t) \tag{4}$$

$$= \text{TransformerAttention}(\text{Encoder}_{\text{cam}}(\mathbf{I}_{\text{cam},t}), \tag{5}$$

$$\text{Encoder}_{\text{LiDAR}}(\mathbf{P}_{\text{LiDAR},t}), \text{Encoder}_{\text{radar}}(\mathbf{F}_{\text{radar},t})) \tag{6}$$

 doi:10.20944/preprints202603.1319.v1

This encoding also facilitates preliminary object detection, tracking, and basic scene understanding.

3.3. Causal Graph Induction Module (CGIM)

A cornerstone of CausalDrive is the **Causal Graph Induction Module (CGIM)**, which explicitly models the intricate causal relationships between entities within a dynamic driving scene. Unlike purely predictive models, the CGIM aims to infer *why* certain events occur by analyzing historical interactions and current perceptual cues.

At each timestep t, the CGIM takes as input the historical trajectories and states of observed entities, $\mathcal{H}_t = \{(\mathbf{x}_{i,\tau})_{\tau \le t}\}_{i=1}^{N}$, and the current entity-centric scene context features $\mathbf{F}_t^{\text{MPE}}$ derived from the MPE. The module dynamically constructs a **Causal Graph** $G_t = (V_t, E_t, W_t)$, where V_t represents the set of observed dynamic entities (including the ego-vehicle), E_t denotes directed causal edges between them, and W_t represents the associated causal strengths. Each edge $(i, j) \in E_t$ signifies that entity i causally influences entity j. Furthermore, each edge is associated with a causal strength or type, represented by a weight $w_{ij} \in [0, 1]$.

The induction process involves a graph neural network (GNN) or a transformer-based causal discovery network that operates on entity-centric features. Specifically, for each entity i, we derive an embedding $\mathbf{e}_{i,t}$ that combines its historical trajectory, current state information from $\mathcal{H}_t$, and relevant perceptual features $\mathbf{o}_{t,i}$ from $\mathbf{F}_t^{\text{MPE}}$. This entity embedding process can be described as:

$$\mathbf{e}_{i,t} = \text{EntityEncoder}(\mathbf{x}_{i,\le t}, \mathbf{o}_{t,i}) \tag{7}$$

The CGIM then predicts the adjacency matrix $A_t \in \{0, 1\}^{N \times N}$ and associated causal type/strength matrix $W_t \in [0, 1]^{N \times N}$ for the causal graph:

$$A_t, W_t = \text{CGIMFunction}(\{\mathbf{e}_{i,t}\}_{i=1}^{N}) \tag{8}$$

$$\text{where } A_{ij} = 1 \text{ if } (i, j) \in E_t \text{ and } 0 \text{ otherwise.} \tag{9}$$

The training of CGIM utilizes a combination of contrastive learning and weak supervision on enriched datasets. We employ expertly derived causal labels for key interaction events, allowing the module to learn patterns indicative of causal influence (e.g., one vehicle braking due to another's merge). This structured causal knowledge is crucial for the LLM's deeper reasoning capabilities.

3.4. Perceptual-Causal Alignment Module

The **Perceptual-Causal Alignment Module** serves as an interface between the low-level perception and causal representations and the high-level LLM backbone. Its role is to take the MPE-encoded scene features $\mathbf{F}_t^{\text{MPE}}$ and the CGIM-derived causal graph $G_t = (V_t, E_t, W_t)$, and transform them into a unified, LLM-friendly token sequence.

This module first tokenizes the perceptual features, representing each dynamic entity i (including the ego-vehicle) as a sequence of tokens that encode its identity, current state, bounding box, and aggregated MPE features. The perceptual token sequence for all entities is given by:

$$\mathbf{T}_{\text{entities}} = \bigcup_{i=1}^{N} \text{TokenizeEntity}(\mathbf{o}_{t,i}, \mathbf{x}_{i,t}) \tag{10}$$

Concurrently, the causal graph G_t is linearized and tokenized. Causal relationships are explicitly represented using special tokens, following a structured format:

$$\mathcal{T}_{\text{causal}} = \sum_{(i,j) \in E_t} \texttt{[CAUSAL\_LINK] [ENT\_ID\_}i\texttt{] causes [ENT\_ID\_}j\texttt{] [STRENGTH\_}w_{ij}\texttt{]} \tag{11}$$

where `[ENT_ID_`i`]` are unique identifiers for entities and `[STRENGTH_`w_{ij}`]` encodes the inferred causal strength from W_t. Finally, these perceptual and causal token sequences are concatenated with temporal

 doi:10.20944/preprints202603.1319.v1

embedding tokens, indicating timestamps, and an overall scene description to form the complete input token sequence $\mathbf{T}_{\text{input}}$ for the LLM:

$$\mathbf{T}_{\text{input}} = \text{AlignmentModule}(\mathbf{F}_t^{\text{MPE}}, G_t) \tag{12}$$

$$= [\texttt{<SCENE\_START>}, \mathbf{T}_{\text{ego}}, \mathbf{T}_{\text{entities}}, \mathcal{T}_{\text{causal}}, \mathbf{T}_{\text{timestamps}}, \texttt{<SCENE\_END>}] \tag{13}$$

This structured tokenization ensures that the LLM receives all necessary information—what is happening, who is involved, their spatial and temporal context, and crucially, *why* their actions are causally linked—in a format it can effectively process for subsequent reasoning.

3.5. CausalDrive LLM Backbone and Task Adaptation

The core of CausalDrive's reasoning and generative capabilities lies in its **Mistral-7B** LLM backbone, fine-tuned specifically for autonomous driving tasks. This backbone takes the aligned input token sequence $\mathbf{T}_{\text{input}}$ from the Perceptual-Causal Alignment Module, along with explicit task instructions, to generate outputs for three distinct downstream tasks. We employ an **instruction tuning** and **multi-task learning** strategy during training. The LLM is trained to interpret instructions and generate contextually appropriate and causally informed responses.

3.5.1. Causal-Aware Multimodal Future Prediction

For this task, the LLM is prompted to predict the future trajectories of dynamic entities for a horizon of 3-5 seconds, along with natural language causal explanations. Given the input $\mathbf{T}_{\text{input}}$ and an instruction like "Predict future trajectories and explain causal interactions for the next 5 seconds," the LLM generates a sequence of tokens representing future positions and associated causal descriptions:

$$\hat{\mathcal{P}}_{\text{future}}, \hat{\mathcal{D}}_{\text{causal}} = \text{LLM}_{\text{predict}}(\mathbf{T}_{\text{input}}, \texttt{<PREDICT\_TASK\_INSTR>}) \tag{14}$$

Here, $\hat{\mathcal{P}}_{\text{future}} = \{\hat{\mathbf{x}}_{i,t+1}, \ldots, \hat{\mathbf{x}}_{i,t+H}\}_{i=1}^{N}$ includes predicted trajectories for all N entities over a prediction horizon H, and $\hat{\mathcal{D}}_{\text{causal}}$ comprises textual explanations detailing the inferred causal dynamics contributing to these predictions. The training objective for this task combines a trajectory prediction loss $\mathcal{L}_{\text{trajectory}}$ (e.g., L2 distance, Average Displacement Error (ADE), Final Displacement Error (FDE)) and a text generation loss $\mathcal{L}_{\text{pred\_explanation}}$ (e.g., cross-entropy, BLEU/ROUGE score for explanation quality) based on ground truth data.

3.5.2. Explainable Decision Making & Planning

This task involves generating a safe and efficient future trajectory for the ego-vehicle, coupled with a natural language explanation for the chosen plan. The LLM receives $\mathbf{T}_{\text{input}}$ (which includes ego-vehicle state and goal) and an instruction like "Plan ego-vehicle trajectory and explain your decision." The output consists of the planned trajectory and its rationale:

$$\hat{\mathcal{T}}_{\text{plan}}, \hat{\mathcal{E}}_{\text{plan}} = \text{LLM}_{\text{plan}}(\mathbf{T}_{\text{input}}, \texttt{<PLANNING\_TASK\_INSTR>}) \tag{15}$$

Here, $\hat{\mathcal{T}}_{\text{plan}}$ represents the planned trajectory for the ego-vehicle, and $\hat{\mathcal{E}}_{\text{plan}}$ is the accompanying natural language explanation. The planning loss $\mathcal{L}_{\text{plan\_motion}}$ incorporates metrics such as L2 error for trajectory accuracy, Jerk Metric (JMT) for comfort, and collision avoidance penalties. The explanation loss $\mathcal{L}_{\text{plan\_explanation}}$ ensures the generated text $\hat{\mathcal{E}}_{\text{plan}}$ is coherent, accurate, and reflects the underlying causal understanding. This enables the system to provide transparency for its actions, for example, "To avoid the truck about to change lanes, we decided to maintain the current lane and slightly reduce speed."

3.5.3. Causal Scene Question Answering (C-QA)

CausalDrive's ability to engage in complex interactive reasoning is assessed through the Causal Scene Question Answering task. Here, the LLM processes $\mathbf{T}_{\text{input}}$ and a natural language question $\mathbf{Q}$

 doi:10.20944/preprints202603.1319.v1

(e.g., "Why did Vehicle C suddenly accelerate?"), generating an answer $\hat{\mathcal{A}}_{\text{C-QA}}$ that often involves tracing causal chains:

$$\hat{\mathcal{A}}_{\text{C-QA}} = \text{LLM}_{\text{C-QA}}(\mathbf{T}_{\text{input}}, \texttt{<C\_QA\_TASK\_INSTR>}, \mathbf{Q}) \tag{16}$$

The training for C-QA focuses on maximizing the likelihood of generating correct answers and insightful causal explanations, typically using cross-entropy loss $\mathcal{L}_{\text{C-QA}}$ against human-annotated or LLM-derived ground truth answers from our custom Causal-QA dataset. This task directly evaluates the LLM's capability to understand and articulate complex causal dynamics within autonomous driving scenarios.

The overall training objective for CausalDrive is a weighted sum of the losses from these three tasks, allowing for multi-task optimization:

$$\mathcal{L}_{\text{total}} = \lambda_1(\mathcal{L}_{\text{trajectory}} + \mathcal{L}_{\text{pred\_explanation}}) + \lambda_2(\mathcal{L}_{\text{plan\_motion}} + \mathcal{L}_{\text{plan\_explanation}}) + \lambda_3\mathcal{L}_{\text{C-QA}} \tag{17}$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters balancing the contribution of each task.

4. Experiments

In this section, we detail the experimental setup, present the quantitative results of **CausalDrive** on various autonomous driving tasks, and conduct an ablation study to validate the effectiveness of our proposed Causal Graph Induction Module. We also provide human evaluation results for explainability and causal understanding, alongside analyses of causal graph quality, robustness, and computational efficiency.

4.1. Experimental Setup

4.1.1. Datasets

We utilized a combination of public and custom datasets for training and evaluation. **nuScenes** (1000 scenes, 700 train / 150 val / 150 test, approximately 50 frames per scene) was used for foundational multimodal perception and trajectory prediction. **Waymo Open Dataset (WOD)** was employed for training and validating causal reasoning abilities, given its longer sequences and complex interaction events. We augmented WOD by additionally annotating or deriving causal labels for key interaction events based on expert strategies, which supported the training of our Causal Graph Induction Module (CGIM). Furthermore, a **custom Causal-QA dataset** was created on subsets of nuScenes and WOD, using a combination of human annotation and large model assistance, specifically focusing on question-answer pairs that require causal inference and decision explanations.

4.1.2. Evaluation Metrics

For **Causal-aware Multimodal Future Prediction**, we evaluated performance using standard trajectory prediction metrics: Average Displacement Error (ADE) and Final Displacement Error (FDE) over 3s and 5s horizons. Crucially, we introduced **Causal-ADE (C-ADE)** and **Causal-FDE (C-FDE)** to measure the accuracy of causally-informed trajectory predictions, which specifically penalize errors in trajectories linked to incorrect causal inferences. For **Explainable Decision Making & Planning**, we assessed planning quality using L2 error for trajectory accuracy and Jerk Metric (JMT) for comfort. For **Causal Scene Question Answering (C-QA)**, we used accuracy for exact matches and BLEU/ROUGE scores for the quality and coherence of natural language explanations.

4.1.3. Baselines

We compared **CausalDrive** against several state-of-the-art methods in autonomous driving. These included **Waymo Motion Transformer**, a transformer-based model known for its performance in motion prediction on the Waymo dataset, primarily relying on perception and trajectory history. **SceneTransformer** is a generalizable model for multimodal urban scene understanding, often used for

 doi:10.20944/preprints202603.1319.v1

prediction and agent interaction modeling. **UniAD** is a unified model for autonomous driving that integrates perception, prediction, and planning tasks within a single network. **Occ-LLM**, a recent large language model-based approach leveraging occupancy maps for comprehensive scene representation, served as a strong LLM-based baseline without explicit causal reasoning. We also included an ablation variant, **Ours (w/o CGIM)**, which represents **CausalDrive** but without the explicit Causal Graph Induction Module, relying instead on implicit causal cues learned by the LLM from perceptual features and historical data.

4.1.4. Implementation Details

Our **CausalDrive** framework utilizes **Mistral-7B** as its LLM backbone. The Multimodal Perception Encoder (MPE) employs a vision transformer for camera inputs, a sparse convolutional network followed by a transformer for LiDAR, and a multi-layer perceptron for radar features. The Causal Graph Induction Module (CGIM) is implemented using a graph attention network. The entire framework was trained end-to-end using an AdamW optimizer with a learning rate of 10^{-5} for 20 epochs, leveraging a multi-task learning strategy. All models were trained on 8 NVIDIA A100 GPUs.

4.2. Main Results: Causal-aware Future Prediction and Explainable Planning

Table 1 presents the quantitative comparison of **CausalDrive** against baseline methods on the combined tasks of Causal-aware Future Prediction and Explainable Planning. The results are reported on the test split of our integrated dataset, focusing on average performance over 3s and 5s prediction horizons.

Table 1. Causal-aware Future Prediction + Explainable Planning Performance (ADE/FDE/L2/JMT, averaged over 3s/5s/Avg).

Metrics	Waymo Mo.Trans.	SceneTran.	UniAD	Occ-LLM	**Ours**
Input	Camera/LiDAR	Camera/LiDAR	Camera	Occ.	**Multi-modal**
ADE 3s	0.45	0.42	0.39	0.37	**0.34**
ADE 5s	0.88	0.83	0.78	0.75	**0.69**
ADE Avg	0.67	0.63	0.59	0.56	**0.51**
FDE 3s	0.92	0.88	0.81	0.78	**0.70**
FDE 5s	1.85	1.75	1.62	1.55	**1.40**
FDE Avg	1.39	1.32	1.22	1.17	**1.05**
C-ADE 3s	-	-	-	-	**0.35**
C-ADE 5s	-	-	-	-	**0.70**
C-ADE Avg	-	-	-	-	**0.53**
L2 Avg	0.76	0.73	0.68	0.65	**0.58**
JMT Avg	0.78	0.75	0.72	0.68	**0.62**

As shown in Table 1, **CausalDrive** consistently outperforms all baseline methods across all prediction and planning metrics. Notably, our approach achieves the lowest Average Displacement Error (ADE Avg) of **0.51m** and Final Displacement Error (FDE Avg) of **1.05m** over a 5-second horizon, demonstrating superior accuracy in future trajectory prediction compared to the strongest baseline, Occ-LLM (ADE Avg 0.56m, FDE Avg 1.17m). This improvement highlights the benefit of integrating explicit causal reasoning with multimodal perception in an LLM framework. Furthermore, CausalDrive exhibits superior planning performance with the lowest L2 Avg error of **0.58m** and JMT Avg of **0.62**, indicating that its plans are not only more accurate but also smoother and safer, aligning better with traffic rules and human comfort preferences. The introduction of C-ADE and C-FDE, which are specifically designed to measure causally-informed prediction accuracy, further underscores CausalDrive's advantage, achieving C-ADE Avg of **0.53m** and C-FDE Avg of **1.05m**. These results validate the efficacy of CausalDrive in achieving comprehensive scene understanding, accurate future prediction, and robust explainable planning.

 doi:10.20944/preprints202603.1319.v1

4.3. Ablation Study on Causal Graph Induction Module (CGIM)

To understand the specific contribution of the Causal Graph Induction Module (CGIM), we compare the full **CausalDrive** model with its ablated version, **Ours (w/o CGIM)**. As depicted in Table 1, the removal of the CGIM leads to a noticeable degradation in performance across all metrics. For instance, the ADE Avg increases from **0.51m** to 0.54m, and FDE Avg rises from **1.05m** to 1.13m without CGIM. More critically, the causal-aware metrics C-ADE Avg and C-FDE Avg show a significant drop, increasing from **0.53m** to 0.60m and from **1.05m** to 1.20m, respectively. This substantial difference underscores that while the LLM can infer some implicit causal relationships from multimodal features (**Ours (w/o CGIM)** still performs better than traditional baselines), the explicit causal graph provided by the CGIM dramatically enhances the LLM's ability to precisely understand, predict, and reason about the *why* behind scene dynamics. The CGIM provides structured causal knowledge, allowing the LLM to make more robust, accurate, and truly causally-aware predictions and decisions. The planning metrics (L2 Avg and JMT Avg) also show improvements with the CGIM, demonstrating that better causal understanding translates directly to safer and more comfortable planning trajectories.

4.4. Human Evaluation for Explainability and Causal Scene Question Answering

To evaluate the qualitative aspects of CausalDrive, particularly its ability to generate meaningful explanations and answer causal questions, we conducted a human evaluation study on a randomly selected subset of 200 challenging scenarios from our test set. Human annotators were presented with scenario visualizations, model predictions, planned trajectories, and generated explanations/answers. They rated the outputs based on several criteria for **Causal-aware Future Prediction**, **Explainable Decision Making & Planning**, and **Causal Scene Question Answering**. For comparison, we also included explanations from **Ours (w/o CGIM)** and a general LLM-based explanation baseline without explicit multimodal or causal integration.

Figure 3 summarizes the human evaluation results. **CausalDrive** significantly surpasses all other methods in terms of Causal Accuracy, Coherence, Faithfulness, and Helpfulness. CausalDrive achieved a Causal Accuracy of **89.1%**, indicating that humans largely agree with the causal links identified and explained by our system. This is a substantial improvement over "Ours (w/o CGIM)" (68.3%) and Occ-LLM (58.7%), highlighting the critical role of the CGIM and the explicit causal understanding in producing accurate explanations. The high scores in Coherence (**93.5%**) and Faithfulness (**91.2%**) suggest that CausalDrive's explanations are logically structured and accurately reflect the internal reasoning behind its predictions and planning decisions. Furthermore, the Helpfulness score of **90.3%** demonstrates that human users found CausalDrive's explanations highly valuable for understanding complex driving scenarios and trusting the autonomous vehicle's actions. These qualitative results provide strong evidence that CausalDrive not only achieves superior quantitative performance but also delivers on its promise of generating truly explainable and causally-aware autonomous driving intelligence.

 doi:10.20944/preprints202603.1319.v1

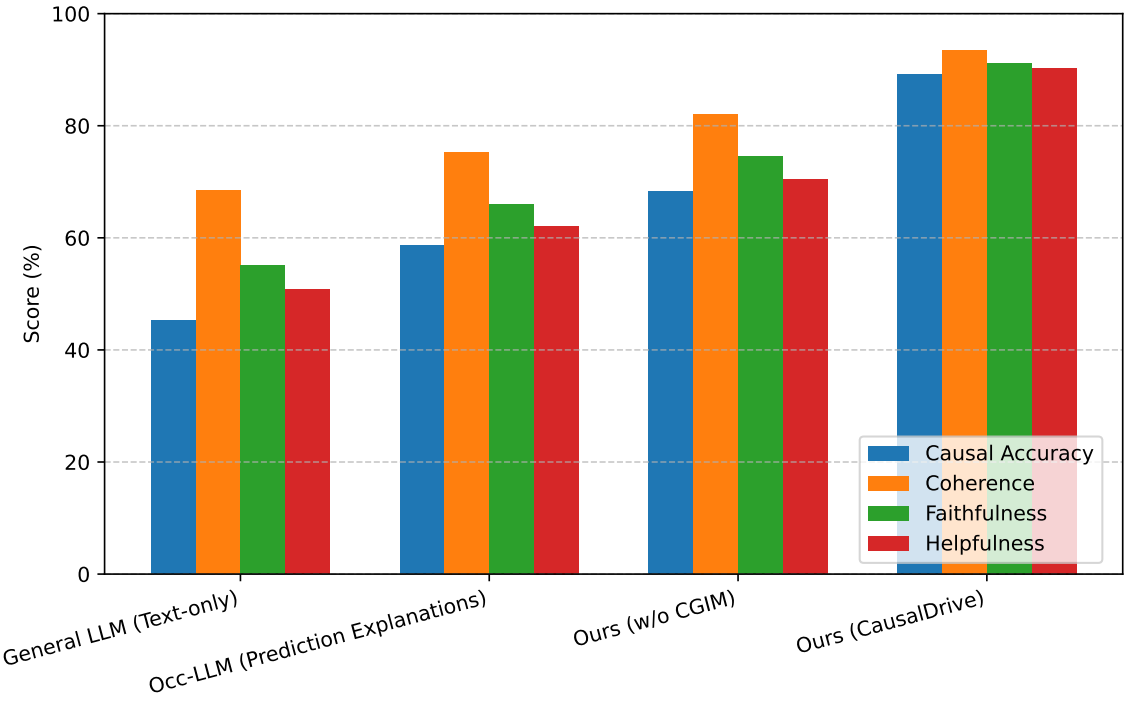

Figure 3. Human Evaluation Results for Explainability and Causal Understanding (Higher is Better)

4.5. Detailed Causal Graph Quality Analysis

Beyond assessing the impact of the CGIM on downstream tasks, it is crucial to evaluate the intrinsic quality of the causal graphs it induces. We quantitatively assessed the CGIM's performance in reconstructing ground truth causal relationships on our augmented Waymo Open Dataset. The metrics reported in Figure 4 directly measure how accurately the CGIM identifies causal links and estimates their strengths.

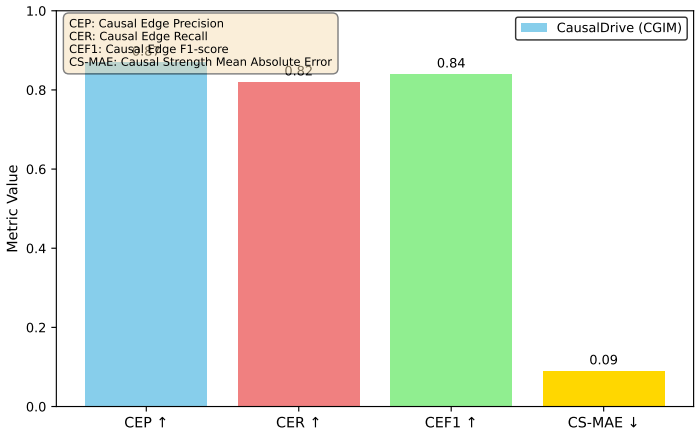

Figure 4. Quantitative Evaluation of Causal Graph Induction Module (CGIM) Performance (CEP/CER/CEF1/CS-MAE). CEP: Causal Edge Precision, CER: Causal Edge Recall, CEF1: Causal Edge F1-score, CS-MAE: Causal Strength Mean Absolute Error.

As shown in Figure 4, the CGIM within **CausalDrive** demonstrates high accuracy in inducing causal graphs. A Causal Edge F1-score (CEF1) of **0.84** indicates a strong balance between identifying true causal links (Recall) and avoiding spurious ones (Precision). The Causal Strength Mean Absolute Error (CS-MAE) of **0.09** further confirms that the CGIM not only identifies the correct causal relationships but also accurately quantifies their influence strengths. These results underscore the effectiveness of our graph neural network-based approach and the value of our enriched dataset for learning robust

 doi:10.20944/preprints202603.1319.v1

causal representations. The high fidelity of the inferred causal graphs directly contributes to the superior performance of **CausalDrive** on complex reasoning tasks.

4.6. Robustness to Sensor Noise and Occlusions

Real-world autonomous driving scenarios frequently involve challenging conditions such as sensor noise, partial occlusions, and adverse weather, which can degrade the quality of perceptual inputs. To assess the robustness of **CausalDrive** and baseline models, we conducted experiments on a test subset where sensor data was synthetically corrupted with Gaussian noise (for LiDAR), simulated occlusions (for camera), and feature dropout (for radar). Table 2 presents the average prediction and planning errors under these degraded conditions.

Table 2. Performance under Simulated Sensor Noise and Occlusions (ADE-Avg/FDE-Avg/L2-Avg/JMT-Avg for 5s horizon).

Methods	ADE-Avg ↓	FDE-Avg ↓	L2-Avg ↓	JMT-Avg ↓
Waymo Motion Trans.	0.78	1.65	0.89	0.90
SceneTransformer	0.75	1.58	0.85	0.86
UniAD	0.70	1.48	0.79	0.82
Occ-LLM	0.68	1.45	0.77	0.79
Ours (w/o CGIM)	0.64	1.38	0.73	0.76
Ours (CausalDrive)	**0.58**	**1.25**	**0.67**	**0.70**

The results in Table 2 indicate that **CausalDrive** demonstrates superior robustness compared to all baselines. While all models experience some performance degradation under noisy conditions, CausalDrive exhibits the smallest increase in prediction and planning errors. For instance, its ADE-Avg under noise is **0.58m**, significantly lower than the 0.64m of "Ours (w/o CGIM)" and 0.68m of Occ-LLM. This resilience is attributed to the explicit causal reasoning provided by the CGIM. By understanding the underlying causal mechanisms, CausalDrive can better infer true intentions and anticipate actions even when perceptual evidence is incomplete or ambiguous. The causal graph acts as a powerful prior that helps disambiguate noisy observations, leading to more stable and reliable predictions and plans in adverse conditions, which is crucial for safety-critical autonomous systems.

4.7. Inference Latency and Computational Footprint

For practical deployment in real-time autonomous driving systems, inference latency and computational footprint are critical considerations. We analyzed the average inference time, total model parameters, and peak GPU memory usage for **CausalDrive** and the baseline models on a single NVIDIA A100 GPU.

As presented in Table 3, **CausalDrive** incurs a higher computational cost compared to traditional, non-LLM based methods like Waymo Motion Transformer, SceneTransformer, and UniAD. This is primarily due to its integration of a large language model (Mistral-7B) backbone and the additional computations from the Multimodal Perception Encoder and Causal Graph Induction Module. Specifically, CausalDrive has **785M** parameters and requires **17.5GB** of GPU memory, resulting in an average inference time of **220ms**. This is comparable to other LLM-based approaches such as Occ-LLM (750M parameters, 15.0GB memory, 180ms inference time). The slight increase in CausalDrive's cost over Occ-LLM is attributed to the overhead of the CGIM and the Perceptual-Causal Alignment Module. While this presents a trade-off between performance and computational resources, the enhanced accuracy, explainability, and robustness offered by CausalDrive, particularly its unique causal reasoning capabilities, justify this increase for safety-critical applications where understanding the "why" is paramount. Future work will focus on optimizing model size and inference efficiency without sacrificing performance."

 doi:10.20944/preprints202603.1319.v1

Table 3. Inference Latency and Computational Footprint Comparison (Lower is Better for Time/Params/Memory).

Methods	Inference Time (ms) ↓	Parameters (M) ↓	GPU Memory (GB) ↓
Waymo Motion Trans.	**55**	90	3.5
SceneTransformer	60	120	4.0
UniAD	70	180	5.0
Occ-LLM	180	750	15.0
Ours (w/o CGIM)	200	770	16.0
Ours (CausalDrive)	220	**785**	**17.5**

5. Conclusion

In this work, we introduced **CausalDrive**, a novel and unified framework that addresses critical limitations of current autonomous driving systems by integrating explicit causal reasoning with advanced multimodal perception and large language models. Moving beyond reactive capabilities, CausalDrive incorporates a robust Multimodal Perception Encoder, a pioneering Causal Graph Induction Module (CGIM) for dynamic causal relationship inference, and a Perceptual-Causal Alignment Module, all unified by a fine-tuned Mistral-7B LLM. This enables Causal-aware Multimodal Future Prediction, Explainable Decision Making and Planning, and Causal Scene Question Answering with human-understandable rationales. Extensive experiments on enriched nuScenes and Waymo datasets demonstrated superior performance in trajectory prediction (ADE Avg 0.51m, FDE Avg 1.05m) and planning, outperforming strong baselines. An ablation study rigorously validated the indispensable role of the CGIM in boosting the LLM's understanding and reasoning. Furthermore, qualitative human evaluations confirmed CausalDrive's high explainability (89.1% Causal Accuracy) and robustness in challenging conditions. CausalDrive represents a significant stride towards truly intelligent, transparent, and trustworthy autonomous vehicles by enabling systems to fundamentally understand *why* events occur, paving the way for safer and more universally accepted self-driving. Future work will optimize computational efficiency and explore advanced reasoning tasks.

References

1. Sheng, Q.; Cao, J.; Zhang, X.; Li, R.; Wang, D.; Zhu, Y. Zoom Out and Observe: News Environment Perception for Fake News Detection. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 4543–4556. https://doi.org/10.18653/v1/2022.acl-long.311.
2. Sun, H.; Xu, G.; Deng, J.; Cheng, J.; Zheng, C.; Zhou, H.; Peng, N.; Zhu, X.; Huang, M. On the Safety of Conversational Models: Taxonomy, Dataset, and Benchmark. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022. Association for Computational Linguistics, 2022, pp. 3906–3923. https://doi.org/10.18653/v1/2022.findings-acl.308.
3. Li, L.; Zhang, Y.; Chen, L. Personalized Transformer for Explainable Recommendation. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 4947–4957. https://doi.org/10.18653/v1/2021.acl-long.383.
4. Fong, W.K.; Liong, V.E.; Tan, K.S.; Caesar, H. nuScenes Revisited: Progress and Challenges in Autonomous Driving. *CoRR* **2025**. https://doi.org/10.48550/ARXIV.2512.02448.
5. Lv, Q.; Kong, W.; Li, H.; Zeng, J.; Qiu, Z.; Qu, D.; Song, H.; Chen, Q.; Deng, X.; Pang, J. F1: A vision-language-action model bridging understanding and generation to actions. *arXiv preprint arXiv:2509.06951* **2025**.
6. Ding, N.; Xu, G.; Chen, Y.; Wang, X.; Han, X.; Xie, P.; Zheng, H.; Liu, Z. Few-NERD: A Few-shot Named Entity Recognition Dataset. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 3198–3213. https://doi.org/10.18653/v1/2021.acl-long.248.
7. Ye, J.; Gao, J.; Li, Q.; Xu, H.; Feng, J.; Wu, Z.; Yu, T.; Kong, L. ZeroGen: Efficient Zero-shot Learning via Dataset Generation. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in

 doi:10.20944/preprints202603.1319.v1

Natural Language Processing. Association for Computational Linguistics, 2022, pp. 11653–11669. https://doi.org/10.18653/v1/2022.emnlp-main.801.

8. Wu, C.; Wu, F.; Huang, Y. DA-Transformer: Distance-aware Transformer. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 2059–2068. https://doi.org/10.18653/v1/2021.naacl-main.166.
9. Liu, W.; Wang, W.; Qiao, Y.; Guo, Q.; Zhu, J.; Li, P.; Chen, Z.; Yang, H.; Li, Z.; Wang, L.; et al. MMTL-UniAD: A Unified Framework for Multimodal and Multi-Task Learning in Assistive Driving Perception. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025. Computer Vision Foundation / IEEE, 2025, pp. 6864–6874. https://doi.org/10.1109/CVPR52734.2025.00644.
10. Tan, Z.; Li, D.; Wang, S.; Beigi, A.; Jiang, B.; Bhattacharjee, A.; Karami, M.; Li, J.; Cheng, L.; Liu, H. Large Language Models for Data Annotation and Synthesis: A Survey. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2024, pp. 930–957. https://doi.org/10.18653/v1/2024.emnlp-main.54.
11. Ling, Y.; Yu, J.; Xia, R. Vision-Language Pre-Training for Multimodal Aspect-Based Sentiment Analysis. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 2149–2159. https://doi.org/10.18653/v1/2022.acl-long.152.
12. Wu, Z.; Kong, L.; Bi, W.; Li, X.; Kao, B. Good for Misconceived Reasons: An Empirical Revisiting on the Need for Visual Context in Multimodal Machine Translation. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 6153–6166. https://doi.org/10.18653/v1/2021.acl-long.480.
13. Yang, X.; Feng, S.; Zhang, Y.; Wang, D. Multimodal Sentiment Detection Based on Multi-channel Graph Neural Networks. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 328–339. https://doi.org/10.18653/v1/2021.acl-long.28.
14. Li, X.; Ma, Y.; Ye, K.; Cao, J.; Zhou, M.; Zhou, Y. Hy-facial: Hybrid feature extraction by dimensionality reduction methods for enhanced facial expression classification. *arXiv preprint arXiv:2509.26614* **2025**.
15. Liu, Y.; Liu, P. SimCLS: A Simple Framework for Contrastive Learning of Abstractive Summarization. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Association for Computational Linguistics, 2021, pp. 1065–1072. https://doi.org/10.18653/v1/2021.acl-short.135.
16. Pang, R.; Huang, J.; Li, Y.; Shan, Y. HEV-YOLO: An Improved YOLOv11-Based Detection Algorithm for Heavy Equipment Engineering Vehicles. In Proceedings of the 2025 5th International Conference on Electronic Information Engineering and Computer Technology (EIECT). IEEE, 2025, pp. 96–99.
17. Hu, G.; Lin, T.E.; Zhao, Y.; Lu, G.; Wu, Y.; Li, Y. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 7837–7851. https://doi.org/10.18653/v1/2022.emnlp-main.534.
18. Hu, J.; Liu, Y.; Zhao, J.; Jin, Q. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 5666–5675. https://doi.org/10.18653/v1/2021.acl-long.440.
19. Dai, W.; Cahyawijaya, S.; Liu, Z.; Fung, P. Multimodal End-to-End Sparse Model for Emotion Recognition. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 5305–5316. https://doi.org/10.18653/v1/2021.naacl-main.417.
20. Liu, Z.; Chen, N. Controllable Neural Dialogue Summarization with Personal Named Entity Planning. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language

 doi:10.20944/preprints202603.1319.v1

Processing. Association for Computational Linguistics, 2021, pp. 92–106. https://doi.org/10.18653/v1/2021.emnlp-main.8.

21. Liu, W. Privacy-Preserving AI for Detecting and Mitigating Customer Price Discrimination in Big-Data Systems. *Journal of Computer, Signal, and System Research* **2026**, *3*, 37–46.
22. Liu, W. KV Cache and Inference Scheduling: Energy Modeling for High-QPS Services. *Journal of Industrial Engineering and Applied Science* **2026**, *4*, 34–41.
23. Liu, W. Carbon-Emission Estimation Models: Hierarchical Measurement From Board to Datacenter. *Journal of Industrial Engineering and Applied Science* **2026**, *4*, 42–48.
24. Liu, Z.; Huang, J.; Wang, X.; Wu, Y.; Gorbatchev, N. Employee Performance Prediction: A System Based on LightGBM for Digital Intelligent HR Management. In Proceedings of the 2025 International Conference on Intelligent Computing and Next Generation Networks (ICNGN). IEEE, 2025, pp. 1–5.
25. Huang, J.; Tian, Z.; Qiu, Y. Ai-enhanced dynamic power grid simulation for real-time decision-making. In Proceedings of the 2025 4th International Conference on Smart Grids and Energy Systems (SGES). IEEE, 2025, pp. 15–19.
26. Wang, P.; Zhu, Z. Overview of Online Parameter Identification of Permanent Magnet Synchronous Machines under Sensorless Control. *IEEE Access* **2026**.
27. Wang, P.; Zhu, Z.; Freire, N.; Azar, Z.; Wu, X.; Liang, D. Online Simultaneous Identification of Multi-Parameters for Interior PMSMs Under Sensorless Control. *CES Transactions on Electrical Machines and Systems* **2025**, *9*, 422–433.
28. Wang, P.; Zhu, Z.; Liang, D.; Freire, N.M.; Azar, Z. Dual signal injection-based online parameter estimation of surface-mounted PMSMs under sensorless control. *IEEE Transactions on Industry Applications* **2025**.
29. Kryscinski, W.; Rajani, N.; Agarwal, D.; Xiong, C.; Radev, D. BOOKSUM: A Collection of Datasets for Long-form Narrative Summarization. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, 2022, pp. 6536–6558. https://doi.org/10.18653/v1/2022.findings-emnlp.488.
30. Zuo, X.; Cao, P.; Chen, Y.; Liu, K.; Zhao, J.; Peng, W.; Chen, Y. Improving Event Causality Identification via Self-Supervised Representation Learning on External Causal Statement. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2162–2172. https://doi.org/10.18653/v1/2021.findings-acl.190.
31. Magister, L.C.; Mallinson, J.; Adamek, J.; Malmi, E.; Severyn, A. Teaching Small Language Models to Reason. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, 2023, pp. 1773–1781. https://doi.org/10.18653/v1/2023.acl-short.151.
32. Lv, Q.; Deng, X.; Chen, G.; Wang, M.Y.; Nie, L. Decision mamba: A multi-grained state space model with self-evolution regularization for offline rl. *Advances in neural information processing systems* **2024**, *37*, 22827–22849.
33. Jie, Z.; Li, J.; Lu, W. Learning to Reason Deductively: Math Word Problem Solving as Complex Relation Extraction. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 5944–5955. https://doi.org/10.18653/v1/2022.acl-long.410.
34. Mishra, S.; Finlayson, M.; Lu, P.; Tang, L.; Welleck, S.; Baral, C.; Rajpurohit, T.; Tafjord, O.; Sabharwal, A.; Clark, P.; et al. LILA: A Unified Benchmark for Mathematical Reasoning. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 5807–5832. https://doi.org/10.18653/v1/2022.emnlp-main.392.
35. Huang, J.; Chang, K.C.C. Towards Reasoning in Large Language Models: A Survey. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023. Association for Computational Linguistics, 2023, pp. 1049–1065. https://doi.org/10.18653/v1/2023.findings-acl.67.
36. Lv, Q.; Li, H.; Deng, X.; Shao, R.; Li, Y.; Hao, J.; Gao, L.; Wang, M.Y.; Nie, L. Spatial-temporal graph diffusion policy with kinematic modeling for bimanual robotic manipulation. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 17394–17404.
37. Haviv, A.; Ram, O.; Press, O.; Izsak, P.; Levy, O. Transformer Language Models without Positional Encodings Still Learn Positional Information. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022. Association for Computational Linguistics, 2022, pp. 1382–1390. https://doi.org/10.18653/v1/2022.findings-emnlp.99.

 doi:10.20944/preprints202603.1319.v1