Research on Multi-modal Data Fusion and Decision Explanation System for Autonomous Driving Based on Large Language Models

Jinyao Lu
University of Nottingham Ningbo China, Ningbo, Zhejiang, China

Abstract: **With the rapid development of autonomous driving technology towards L4 (highly automated), public trust in unmanned driving systems has become increasingly prominent. The "decision black box" characteristic of current autonomous driving systems makes it difficult for users to understand their decision logic, which has become a key obstacle restricting technology promotion. This paper proposes a multi-modal data fusion and decision explanation system based on large language models, aiming to improve the transparency and interpretability of autonomous driving systems. The system adopts the Transformer architecture, integrates multi-modal data such as LiDAR point clouds, camera images, and text instructions, achieves semantic-level understanding through feature-level fusion, and generates natural language decision explanations. This research provides a new solution for explainable artificial intelligence in autonomous driving systems, with significant theoretical value and practical significance.**

Keywords: Autonomous Driving; Large Language Models; Multi-Modal Data Fusion; Explainable AI; Decision Explanation

1. Introduction

As a frontier field deeply integrating artificial intelligence and the automotive industry, autonomous driving technology is rapidly advancing towards L4 high automation. Although relevant research reports indicate significant public expectations for unmanned driving safety, survey data from organizations like Survey USA reveals another reality: a considerable portion of the public considers current unmanned driving technology still "immature" [1]. The core issue undermining public trust lies in the "decision black box" characteristic of autonomous driving systems-users cannot understand why the vehicle makes specific decisions, which has become a key obstacle limiting technology adoption [2].

Current research has made significant progress in multi-modal perception fusion and decision technologies, with early, mid-level, and late fusion strategies each having their own advantages and disadvantages [3]. Large language models have also demonstrated strong potential in cross-modal tasks such as visual question answering and image captioning [4]. However, existing research still exhibits notable shortcomings in real-time explanation generation for dynamic driving environments, end-to-end integration of multi-modal fusion and explanation, and balancing system interpretability with performance [5]. Addressing these research gaps, this paper proposes a multi-modal data fusion and decision explanation system for autonomous driving based on large language models, aiming to break the decision black box through natural language explanations, thereby enhancing system transparency and public trust. The research covers key technologies such as spatio-temporal alignment of multi-modal data, Transformer-based feature fusion, and real-time explanation generation [6, 7], offering innovative value at the methodological, technical implementation, and application scenario levels.

2. Related Theories and Technical Foundations

2.1 Overview of Autonomous Driving Multi-modal Data

Autonomous driving systems rely on various sensors to acquire environmental information. The main data modalities include [8]:

(1) Visual Data (Camera)

Visual data, obtained via cameras, contains rich texture, color, and semantic information. Its characteristics include:

• Data Format: RGB images, typically with

resolutions of 1920×1080 or higher.
• Sampling Frequency: 30-60 FPS.
• Information Dimension: 2D spatial information; 3D information can be reconstructed using multiple cameras.
(2) Point Cloud Data (LiDAR)
LiDAR provides precise three-dimensional spatial information with the following characteristics:
• Data Format: Point cloud data, where each point contains (x, y, z, intensity).
• Sampling Density: 64-beam or 128-beam LiDAR.
• Measurement Accuracy: Centimeter-level precision, with a maximum range of 100-200 meters.
(3) Text and Voice Instructions
Natural language instructions provide high-level semantic guidance for autonomous driving systems:
• Instruction Types: Navigation commands, behavior commands, query commands.
• Input Methods: Voice recognition, text input, gesture recognition.
• Semantic Complexity: Ranges from simple commands to complex intent expressions.

Multi-modal Data Characteristics Analysis

Autonomous driving multi-modal data possesses the following key characteristics:
Spatio-temporal Consistency Requirements:
• Synchronization time thresholds.
Heterogeneity Challenges:
• Data Format Differences: Images (matrix), point clouds (point sets), text (sequences).
• Different Sampling Frequencies: Camera 30Hz, LiDAR 10Hz, GPS 1Hz.
• Inconsistent Information Dimensions: 2D vs. 3D, dense vs. sparse.

2.2 Technical Principles of Large Language Models

Core Components of Transformer Architecture

The Transformer architecture is based on the self-attention mechanism [9].

Key Technologies of Large Language Models

(1) Pre-training Strategies
(2) Instruction Fine-tuning

2.3 Multi-modal Data Fusion Methods

Fusion Level Classification

(1) Early Fusion (Data-level Fusion)
(2) Mid-level Fusion (Feature-level Fusion)
(3) Late Fusion (Decision-level Fusion)

Fusion Strategy Comparison

A comparison of different fusion strategies is presented in Table 1.

Table 1. Comparison of Multi-Modal Fusion Strategies

Fusion Level	Advantages	Disadvantages	Applicable Scenarios
Early Fusion	Low information loss,full utilization of raw data	Difficult data alignment,high computational load	Strong sensor complementarity
Mid-level Fusion	High flexibility,can handle heterogeneous features	Requires designing complex fusion networks	Inconsistent feature extraction quality
Late Fusion	Simple and efficient,independent processing per modality	Insufficient information fusion	Balanced decision-making capability per modality

3. System Design and Methodology

3.1 Overall System Design

Design Goals and Principles

The proposed Multi-modal Autonomous Driving System with Explanation (MADSE) follows these design principles:
Design Goals:
1) High-precision Fusion: Multi-modal data fusion accuracy > 90%.
2) Real-time Performance: End-to-end latency < 200ms.
3) Interpretability: BLEU score for generated natural language explanations > 0.75.
4) Robustness: Performance degradation in adverse weather < 15%.
Design Principles:
1) Modular Design: Decoupled functional modules for easy maintenance and upgrading.
2) Scalability: Supports rapid integration of new sensors and modalities.
3) Safety Priority: Transparent decision process for easy verification and auditing.

3.2 Overall System Architecture

The system adopts a layered architecture design:
1) Application Layer: User interaction interface.
2) Decision Explanation Layer: Natural language explanation generation.
3) Fusion Decision Layer: Multi-modal fusion and decision-making.
4) Feature Extraction Layer: Feature extraction and encoding for each modality.

5) Data Acquisition Layer: Sensor data acquisition and preprocessing.

The system's mathematical model is described as follows:

Let the system state S_t at time t be represented as:

$$S_t = \{V_t, L_t, T_t, C_t\}$$

where:

- V_t: *Visual feature vector.*
- L_t: *LiDAR feature vector.*
- T_t: *Text instruction feature vector.*
- C_t: *Context feature vector.*

The system decision function D is defined as:

$$D(S_t) = Softmax(W \cdot F(S_t) + b)$$

here F is the multi-modal fusion function, and W and b are learnable parameters.

The system implementation is based on a defined technology stack.

4. Multi-Modal Data Preprocessing Module

4.1 Spatio-Temporal Alignment Algorithm

Spatio-temporal alignment is fundamental for multi-modal fusion and requires solving the following mathematical problem:

Problem Definition:

Given m modality data streams $\{ D_i(t_i) \}_{i=1}^m$, each with its own timestamp t_i, the goal is to find mapping functions f_i such that:

$$D^i(t) = f_i(D_i(t_i)), s.t. |t - t_i| < \varepsilon$$

Solution:

1) Time Synchronization: Based on hardware timestamps or NTP protocol.

2) Spatial Registration: Transforming data from each modality to a unified coordinate system using calibration parameters.

The system adopts a hybrid fusion strategy, combining the advantages of early, mid-level, and late fusion for data preprocessing.

4.2 Real-Time Processing Optimization

Computational Complexity Analysis

The computational complexity of the fusion system primarily stems from:

1. Attention Calculation: $O(N^2 \cdot d)$, where N is the sequence length and d is the feature dimension.

2. Feature Transformation: $O(m \cdot d^2)$, where m is the number of modalities.

3. Fusion Operation: $O(L \cdot d^2)$, where L is the number of fusion layers.

5. LLM-based Decision Explanation Generation

5.1 LLM Model Selection and Adaptation

Large Language Models (LLMs) are neural network-based AI models focused on natural language processing tasks.

LLM Model Selection Principles and Strategies

For the autonomous driving decision explanation generation task, LLM selection requires comprehensive consideration of computational resources, real-time requirements, Chinese language support capability, and task characteristics [2, 4]. This paper establishes a multi-dimensional evaluation matrix and employs a quantitative scoring system to select the most suitable model.

Selection Evaluation Index System

Let the model evaluation function be:

$$S(M)=\alpha \cdot P\_perf+\beta \cdot P\_latency+\gamma \cdot P\_memory+\delta \cdot P\_chinese$$

where:

- P_perf: *Explanation generation performance score.*
- $P\_latency$: *Inference latency score.*
- P_memory: *Memory requirement score.*
- $P\_chinese$: *Chinese support capability score.*
- α, β, γ, δ: *Weight coefficients, satisfying* $\alpha + \beta + \gamma + \delta = 1$

Scenario-based Selection Strategy

Based on deployment environment and requirement differences, a hierarchical selection strategy is formulated:

1) Edge Device Deployment Scenario

- Main Constraints: Memory limitation (<16GB), real-time requirement (<300ms).
- Computational Complexity: $O(n \cdot d^2)$, where n is sequence length, d is hidden dimension.

2) Cloud High-performance Deployment Scenario

- Main Constraints: Explanation quality, low latency (<100ms).
- Cost Function: $C = \lambda \cdot \text{API\_cost} + \mu \cdot \text{latency}$.

3) Balanced Deployment Scenario

- Main Constraints: Cost-performance balance.
- Optimization Goal: $\min(\text{cost}) \quad \text{s.t.} \quad \text{latency} < \tau, \text{accuracy} > \theta$

Model Adaptation and Fine-tuning Methods

Fine-tuning Strategy Design

A hierarchical fine-tuning strategy is designed for different data scales and task requirements:

1) Few-shot Fine-tuning

• Applicable Condition: Labeled data < 1000 samples.

• Fine-tuned Parameters: Only the adaptation layer + the last 2-3 Transformer layers.

•Loss Functio :

$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{KL}}$.

2) Parameter-Efficient Fine-tuning

• Method: LoRA (Low-Rank Adaptation).

• Formula:

$W' = W + \Delta W = W + BA$,

where $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$, $r \ll \min(d, k)$.

• Advantage: Only requires fine-tuning about 0.1% of parameters, significantly reducing computational demand.

3) Full Fine-tuning

• Applicable Condition: Large-scale labeled data > 10,000 samples).

• Training Objective: Maximize the log-likelihood of explanation generation.

• Optimization Function:

$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log P(w_t \mid w_{<t}, F, \theta)$.

1) Computational Resource Optimization

To meet real-time requirements, the following optimization techniques are adopted:

Quantization Compression

• Method: INT8 quantization, converting 32-bit floating-point numbers to 8-bit integers.

• Compression Ratio:

$\text{compression\_ratio} = \frac{32}{8} = 4\times$

• Accuracy Loss: < 1% (optimized via calibration).

2) Attention Optimization

• Adoption of Flash Attention algorithm.

• Complexity reduced from $O(n^2)$ to $O(n)$

• Memory usage reduced by 50%.

3. Batch Processing Optimization

• Dynamic batching strategy.

• Formula:

$B_{\text{optimal}} = \min\left(\frac{M_{\text{available}}}{M_{\text{per\_sample}}}, B_{\text{max}}\right)$

5.2 Prompt Engineering and Explanation Template Design

Prompt Engineering Methodology

Prompt engineering is the key bridge connecting multi-modal features with natural language explanations. This paper proposes a structured prompt design framework containing four core components:

$$Prompt = T(Role, Context, Task, Format)$$

where:

• T: *Template function.*

• *Role*: *System role definition.*

• *Context*: Specific task description

• *Task*: *Specific task description.*

• *Format*: Output format requirements.

Structured prompt design can effectively guide LLM output, a technique beneficial for explanation generation [4, 8].

Example Structured JSON:

```
json
{
  "decision": "Emergency braking",
  "reason": "Detected a pedestrian suddenly crossing ahead",
  "confidence": 0.95,
  "timestamp": "2024-01-15T14:30:25Z",
  "sensor_data": {
    "camera": "Recognized pedestrian silhouette",
    "lidar": "Distance 8 meters, relative speed 5 km/h"
  }
}}
```

5.3 Real-time Explanation Generation Algorithm

Real-time Generation Architecture Design

To meet the stringent real-time requirements of autonomous driving, a streaming explanation generation architecture is designed.

Streaming Generation Algorithm

To reduce generation latency, prefix caching and streaming decoding strategies are employed.

Latency Optimization Techniques

Multiple optimization techniques are used to meet real-time requirements:

Incremental Generation Technique

• Method: Decompose explanation into multiple segments, output while generating.

• First-token latency $T_{\text{first\_token}} < 100 \text{ ms}$.

• Formula:

$T_{\text{total}} = T_{\text{first\_token}} + (n - 1) \cdot T_{\text{token}}$

2. Model Distillation Optimization

• Student Model: Distilled Qwen-7B (high efficiency).

• Distillation Loss:

\mathcal{L}{\text{KD}} =
\alpha \mathcal{L}{\text{CE}} + .
(1 − \alpha) \mathcal{L}_{\text{KL}}
3. Computational Graph Optimization
• Operator Fusion: Merge multiple small operators into a larger one.
• Memory Optimization: Reduce intermediate variable storage.
• Parallel Computing: Utilize GPU multi-core parallel processing.
Caching and Pre-computation Strategies
A multi-level caching architecture is designed to improve explanation generation efficiency.
Cache Level Design
1) L1 Cache (Feature-level Cache)
oStorage: Explanation templates for common feature combinations.
oHit Rate: ~60%.
oAccess Time: <5 ms.
2) L2 Cache (Scenario-level Cache)
oStorage: Explanations for complete scenarios.
oHit Rate: ~30%.
oAccess Time: <10 ms.
3) L3 Cache (User-level Cache)
oStorage: Personalized explanation preferences.
oHit Rate: ~10%.
oAccess Time: <20 ms.
Trade-off Optimization Between Quality and Efficiency
Under real-time constraints, a balance between explanation quality and generation efficiency must be achieved. This paper proposes an adaptive quality adjustment mechanism.
Dynamic Quality Adjustment
Generation parameters are dynamically adjusted based on system load and user requirements.

6. Conclusion

This paper addresses the "decision black box" problem in autonomous driving systems by proposing a Multi-modal Data Fusion and Decision Explanation System (MADSE) based on large language models. By fusing multi-modal information such as LiDAR, camera, and text instructions, and utilizing the Transformer architecture for feature-level semantic understanding, the system can generate natural language decision explanations that align with human cognition, effectively enhancing the transparency and interactivity of autonomous driving.

Research indicates that the proposed fusion and explanation generation method achieves high explanation quality and robustness while ensuring real-time performance (end-to-end latency < 200ms). Through mechanisms such as structured prompt engineering, streaming generation, and multi-layer caching, the system realizes efficient and comprehensible decision outputs in dynamic driving scenarios.

This study provides an end-to-end solution for explainable AI in autonomous driving. It offers methodological innovation in the integration of multi-modal fusion and natural language explanation, and provides a technical path and empirical reference for building trustworthy and transparent autonomous driving systems. Future work will further explore validation in complex extreme scenarios and personalized explanation generation, promoting the development of autonomous driving technology towards greater safety and reliability.

References

[1] Survey USA. (2024). Public Perception of Autonomous Driving Technology. *Survey USA Reports*.

[2] Zhang, Q., Gui, T., Zheng, R., & Huang, X. (2023). Large-scale language models: From theory to practice. Beijing, China: Publishing House of Electronics Industry.

[3] Qiu, X. (2019). Neural networks and deep learning. Beijing, China: China Machine Press.

[4] Hao Shao, Yuxuan Hu, Letian Wang, Steven L. Waslander, Yu Liu, Hongsheng Li (2023). LMDrive: Closed-loop end-to-end driving with large language models. *Proceedings of the IEEE International Conference on Computer Vision*, 1-10.

[5] Yuan, J., Sun, S., Omeiza, D., Zhao, B., Newman, P., Kunze, L., & Gadd, M. (2024). RAG-Driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model.

[6] Junyi Ma, Xieyuanli Chen, Jiawei Huang, Jingyi Xu, Zhen Luo, Jintao Xu, Weihao Gu, Rui Ai, Hesheng Wang (2023). “Cam4DOcc: Benchmark for camera-only 4D occupancy forecasting in autonomous driving applications”. Shanghai Jiao Tong University.

[7] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, Jiwen Lu (2023). SelfOcc: Self-supervised vision-based 3D occupancy prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition, 1-12.

[8] Walter Zimmer, Gerhard Arya Wardana, Xingcheng Zhou, Rui Song, Suren Sritharan, Alois C. Knoll (2023). TraffiX-A V2X dataset for multi-modal cooperative 3D object detection of traffic participants using onboard and roadside sensors. *IEEE Transactions on Intelligent Transportation Systems*, 24(8), 1-15.

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[10] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.

