運転判断の因果関係を考慮した自動運転のための言語説明データセットの構築

EP22071 佐藤克昭　　指導教授：山下隆義

1. はじめに

視覚情報と言語知識を統合した Multimodal Large Language Model (MLLM) を用いた自動運転手法は，運転判断の根拠を言語として表現可能な手法として注目されている．MLLM の学習には，映像に対する判断とともに，その判断の根拠に対応する言語キャプションを大量に用いた学習が必要となる．既存のデータセットの言語キャプションには，道路構造や周辺物体の状態などの運転判断に関与する外界情報が多く含まれる．しかし，運転判断に至る過程を構成要素として明示的に表現することが経路予測精度にどのような影響を与えるのかについては，明らかになっていない．本研究では，運転判断に至る推論過程を自然言語として明示的に付与したデータセットを構築し，各説明要素の有無が経路予測精度に与える影響を比較する．

2. MLLM を用いた End-to-End 自動運転

MLLM を用いた自動運転の代表的なアプローチとして EMMA [1] が提案されている．EMMA は，複数のカメラで撮影した全周囲の画像と自車両の走行履歴およびナビゲーション指示を自然言語で入力する．視覚情報と言語情報を MLLM で共通の特徴表現へ変換することで，経路計画などの自動運転タスクを，タスク固有のプロンプトに基づき統一的に処理可能である．しかし，このような説明可能な自動運転を実現するためには，運転判断の根拠を明示的に含んだデータセットが必要となる．既存の自動運転データセットは，知覚・運動タスクを中心に構成されており，運転判断に至る因果関係が明示的に記述されていない課題がある．

3. 提案手法

本研究は，運転判断に至る推論過程を 3 つの説明タスクとして定義する．マルチセンサ情報を収録した nuScenes[2] に運転判断に関する自然言語記述を付与することで，推論過程を段階的に表現した自動運転データセットを構築する．データセット構築のフレームワークの概要を図 1 に示す．

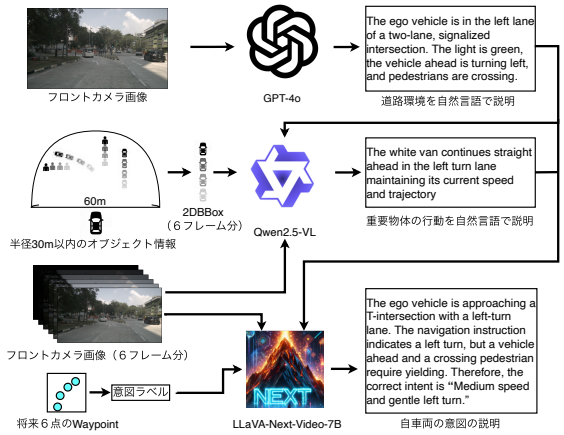

図 1: データセット構築のフレームワーク

道路構造の記述

フロントカメラ画像から，交差点，車線区分，信号，自車両位置などを言語キャプションで生成する．生成には空間理解能力に優れる GPT-4o を用いる．

重要物体の検出および将来行動説明

nuScenes に収録されている画像内の各オブジェクトの BBox，クラスラベルや属性情報を用いる．将来 3 秒間の自車両経路から半径 30 m 以内に存在する物体を重要物体と定義し，道路構造の説明を補助情報として与え，これらの重要物体の状態および将来行動を言語キャプションで生成する．生成には Qwen2.5-VL-7B-Instruct を用いる．

自車両の運転意図の説明

速度および軌道形状をもとに 16 種類の運転意図にルールベースで分類する．さらに，道路構造，重要物体の将来行動およびナビゲーション指示を根拠として，運転意図に対する判断理由を生成する．生成には LLaVA-Next-Video-7B を用いる．これらの説明要素を MLLM に推論させることで，経路予測精度の向上を図る．

4. 評価実験

提案手法の有効性を検証するため，評価実験を行う．LLaVA-Next-Video-7B をベースモデルとし，学習率 1e-5，エポック数 5 で学習を行う．比較手法として，追加学習を行わずに推論を行わせるプロンプトチューニングをベースラインとして用いる．評価指標は経路予測における平均 L2 誤差を用いる．また，各説明要素の有無を変化させ，それらが経路予測精度に与える影響を比較する．

4.1. 定量的評価

表 1 に各説明要素の有無による L2 誤差を示す．結果から，プロンプトチューニングのみでは，「意図説明＋経路予測」と比較して経路予測精度が低い傾向が見られた．また，「経路予測のみ」と比較すると，説明要素の付与により経路予測精度が向上し，特に「意図説明＋経路予測」で顕著な改善が確認された．これは，意図説明が意図ラベルに加えて道路構造および重要物体の将来行動を統合した表現であり，経路生成に有効に作用したためと考えられる．一方，「重要物体＋経路予測」のみの場合は，半径 30 m 以内の物体を重要物体と定義しているため，経路予測に直接関与しない物体が含まれ，精度が低下したと考えられる．

表 1: 説明要素の有無による平均 L2 誤差の比較

提案手法				平均 L2 誤差 [m] ↓
道路構造	重要物体	意図説明	経路予測	
			✓	2.45
✓			✓	2.29
	✓		✓	2.49
✓	✓		✓	2.37
		✓	✓	**2.26**
プロンプトチューニング				4.96

4.2. 定性的評価

図 2 に比較結果を示す．「経路予測のみ」では，横断歩行者が存在する状況においても直進する経路が生成されている．一方，「意図説明＋経路予測」では，横断歩行者を認識した上で「停止」という意図を選択しており，状況認識と運転行動の因果関係が明確に表現されている．

図 2: 説明要素が寄与した一例

5. おわりに

本研究では，自動運転タスクに特化したマルチタスクデータセットを構築し，有効性を検証した．その結果，説明要素の活用により経路予測精度が向上することが判明した．今後は，ベースモデルの変更や異なるドメインのデータセットでの検証を行うとともに，交通ルールを収録した Retrieval Augmented Generation を用いて未学習の地域の交通ルールに対応可能な手法を目指す．

参考文献

[1] J.-J. Hwang, et al., “EMMA: End-to-End Multimodal Model for Autonomous Driving”, arXiv preprint arXiv:2410.23262, 2024.

[2] H. Caesar, et al.,“nuScenes: A Multimodal Dataset for Autonomous Driving”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.