EMMA: End-to-End Multimodal Model for Autonomous Driving

Jyh-Jing Hwang[*] [†], **Runsheng Xu**[*], **Hubert Lin**[‡], **Wei-Chih Hung**[‡], **Jingwei Ji**, **Kristy Choi**, **Di Huang**, **Tong He**, **Paul Covington**, **Benjamin Sapp**, **James Guo**, **Dragomir Anguelov**, **Mingxing Tan**[†]

Waymo LLC

Abstract

We introduce EMMA, an End-to-end Multimodal Model for Autonomous driving. Built on a multi-modal large language model foundation, EMMA directly maps raw camera sensor data into various driving-specific outputs, including planner trajectories, perception objects, and road graph elements. EMMA maximizes the utility of world knowledge from the pre-trained large language models, by representing all non-sensor inputs (e.g. navigation instructions and ego vehicle status) and outputs (e.g. trajectories and 3D locations) as natural language text. This approach allows EMMA to jointly process various driving tasks in a unified language space, and generate the outputs for each task using task-specific prompts. Empirically, we demonstrate EMMA's effectiveness by achieving state-of-the-art performance in motion planning on nuScenes as well as competitive results on the Waymo Open Motion Dataset (WOMD). EMMA also yields competitive results for camera-primary 3D object detection on the Waymo Open Dataset (WOD). We show that co-training EMMA with planner trajectories, object detection, and road graph tasks yields improvements across all three domains, highlighting EMMA's potential as a generalist model for autonomous driving applications. However, EMMA also exhibits certain limitations: it can process only a small amount of image frames, does not incorporate accurate 3D sensing modalities like LiDAR or radar and is computationally expensive. We hope that our results will inspire further research to mitigate these issues and to further evolve the state of the art in autonomous driving model architectures.

1 Introduction

Autonomous driving technology has made significant progress in recent years. To make autonomous vehicles a ubiquitous form of transportation, they must navigate increasingly complex real-world scenarios that require understanding rich scene context as well as sophisticated reasoning and decision-making.

Historically, autonomous driving systems employed a modular approach, consisting of specialized components for perception [Yurtsever et al., 2020, Li et al., 2022b, Lang et al., 2019, Sun et al., 2022, Hwang et al., 2022], mapping [Li et al., 2022a, Tancik et al., 2022], prediction [Nayakanti et al., 2023, Shi et al., 2024], and planning [Teng et al., 2023]. While this design lends itself to easier debugging and optimization of individual modules, it poses scalability challenges due to the accumulated errors among modules and limited inter-module communication. In particular, the expert-designed interfaces between modules, such as the perception and behavior modules, may struggle to adapt to

[*] Equal contributions; [‡] Equal contributions.

[†] Contact emails: Mingxing Tan <tanmingxing@waymo.com>, Jyh-Jing Hwang <jyhh@waymo.com>.

Technical report. October 30, 2024.