ABSTRACT

Autonomous driving systems are undergoing a paradigm shift, transitioning from modular pipelines and end-to-end learning frameworks towards Foundation Model-Driven Autonomy (FMAD). While foundation models, particularly Large Language Models (LLMs) and Vision-Language Models (VLMs), offer unprecedented reasoning capabilities, their deployment is hindered by a Triangle of Limitations: the robustness gap in holistic spatio-temporal perception, the efficiency and alignment gap in multimodal understanding, and the adaptability gap in language-driven planning.

This dissertation proposes the FMAD system, a unified framework designed to bridge these gaps through three complementary layers. First, to establish a robust sensory foundation, we introduce the Driver Digital Twin (DDT) to capture the in-cabin human state, ensuring a human-centric understanding of the driving context. Complementing this, we present CEMFormer, a cross-view episodic memory transformer that captures long-term temporal dependencies for intention prediction, and MACP (Efficient Model Adaptation for Cooperative Perception), a framework that leverages vehicle-to-vehicle (V2V) collaboration to resolve occlusions while minimizing communication costs.

Second, to address the efficiency and alignment challenges, we develop Video Token Sparsification (VTS), a dynamic pruning technique that accelerates multimodal LLM inference by removing redundant visual information. We further propose Multimodal Task Alignment (MTA), which harmonizes Bird's-Eye-View (BEV) perception with dense captioning to reduce hallucinations and enforce semantic consistency.

Finally, within the cognitive layer, this research integrates LLMs into the decision-making loop. We introduce LaMPilot, a framework that translates natural language instructions into executable driving policies, validated by the LaMPilot-Bench. This is further advanced by a human-in-the-loop programming planner, which continuously refines driving behaviors based on human feedback. Collectively, these contributions pave the way for the next generation of autonomous agents that are perceptually robust, computationally efficient, and cognitively adaptive.