ABSTRACT

Autonomous driving is experiencing a paradigm shift from solely pursuing technical autonomy to building systems that align with human intentions, preferences, and trust. However, traditional autonomous vehicles are still designed as automated "tools" which are optimized for perception, planning, and control, however lack the ability to interpret natural human communication or adapt to personalized driving styles. This mismatch hampers user acceptance and limits the deployment of autonomy.

This dissertation introduces Human-Autonomy Teaming (HAT) as a new paradigm and demonstrates that recent advances in Foundation Models, including Large Language Models (LLMs) and Vision Language Models (VLMs), provide the missing capabilities needed to realize this paradigm in autonomous driving. This work systematically proposes, develops, benchmarks, and validates this new paradigm.

The dissertation begins by formalizing the conceptual framework of HAT for autonomous driving, integrating autonomy foundations, human foundations, core teaming processes, and closed-loop feedback. To situate HAT within broader research progress, the dissertation provides the first unified survey of Multimodal Large Language Models (MLLMs) for autonomous driving, covering perception, reasoning, planning, control, benchmarks, and emerging industry applications.

Building on the human foundation, the work introduces VLM-DC, a semantic grounding and data collection ecosystem that aligns scene understanding and data collection with human relevance. Through two different architectures, the system identifies semantically meaningful driving scenes, reduces redundancy, and constructs personalized data distributions for different kinds of vehicles that support human-aligned autonomy.

On the autonomy side, the dissertation presents ViLaD, a vision-language diffusion generative planning model that produces both quick and accurate driving decisions. By replacing autoregressive generation with parallel masked denoising, ViLaD improves robustness, inference efficiency, and the expressiveness needed to support collaborative decision processes.

The dissertation then validates the HAT paradigm through teaming behaviors in simulation. The Receive, Reason, React (RRR) framework demonstrates that LLMs can serve

as the core teaming engine in simulation, interpreting natural language instructions and generating context-aware driving policies. To evaluate such agents, the dissertation introduces LaMPilot, the first benchmark designed for instruction-following and reasoning in autonomous driving through a "code as policy" strategy.

Building on these foundations, the dissertation conducts extensive real-world field experiments to demonstrate HAT-enabled autonomy. The Talk2Drive framework shows that cloud-based LLMs can interpret diverse natural language commands, reason over driving context, and personalize vehicle behavior through memory, significantly reducing driver takeovers. To overcome network-latency limitations, this work further presents the first on-board VLM-based personalized motion control system, achieving low-latency collaboration with a RAG-enhanced memory module and reducing takeovers by 76.9% in real driving tests.

Together, these contributions present a comprehensive demonstration of moving pure automation toward true HAT, providing conceptual foundations, comprehensive reviews, technical foundations, simulation verification, and real-world validation. This dissertation demonstrates how foundation models can enable autonomous vehicles that communicate, reason, adapt, and collaborate, building the next generation of safe, trustworthy, and personalized autonomous vehicles.