LLM-Raft: Enhancing Urban Traffic Efficiency and Safety through Decentralized Coordination of Autonomous Vehicles

Lingfeng Zhou* **Shuaixing Chen*** **Jin Gao** **Dequan Wang**
Shanghai Jiao Tong University
{zhoulingfeng, alkdischen, gaojin, dequanwang}@sjtu.edu.cn

Abstract

Urban areas face persistent challenges of traffic congestion and safety, which hinder efficiency and quality of life. Coordinated autonomous vehicles (AVs) offer a promising solution, but achieving robust, decentralized coordination in dynamic urban settings remains a significant hurdle. This paper introduces LLM-Raft, a novel framework designed to enhance urban mobility by enabling LLM-powered AVs to coordinate their actions safely and efficiently. Inspired by the Raft algorithm, LLM-Raft allows vehicles to generate and agree upon "traffic narratives"—human-like, structured propositions of their intent and justification. This semantic consensus mechanism allows for more intelligent and predictable group behaviors without a central coordinator. We validate our framework in realistic urban traffic simulations. The results show that LLM-Raft improves key urban mobility metrics, reducing collision rates by 40-50% and task completion time by 20-30% compared to uncoordinated baselines. Our work presents a viable path toward more collaborative and resilient multi-agent systems, contributing to the development of safer and more efficient urban transportation networks. Code is available at `https://github.com/shxingch/llm-raft`.

1 Introduction

Urban areas are the engines of the global economy, but they face growing pressure from traffic congestion and safety issues. Inefficient traffic flow not only leads to significant economic losses and environmental pollution but also diminishes the quality of life for city residents. As cities become denser, the need for intelligent transportation systems that can improve efficiency and safety is more urgent than ever. Coordinated autonomous vehicles (AVs) are widely seen as a key technology to address these urban challenges, promising to create smoother and safer traffic patterns.

However, simply deploying a large number of autonomous vehicles is not enough. Without effective coordination, independent AVs might act cautiously and inefficiently, potentially worsening traffic congestion rather than alleviating it. A centralized coordinator could impose a global plan, but this approach creates a single point of failure and a communication bottleneck [29], making it unsuitable for the dynamic and large-scale nature of urban traffic. Therefore, a robust, scalable, and decentralized coordination mechanism is essential for AVs to realize their full potential in urban environments.

The emergence of Large Language Models (LLMs) in autonomous driving presents a major opportunity to make vehicle coordination far more intelligent. With LLMs, individual vehicles can now reason about complex traffic situations and form high-level plans, much like human drivers [26, 27, 21, 24, 1]. However, traditional consensus algorithms [4, 33, 5] typically synchronize low-level kinematic data,

*Equal contribution.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: UrbanAI: Harnessing Artificial Intelligence for Smart Cities.

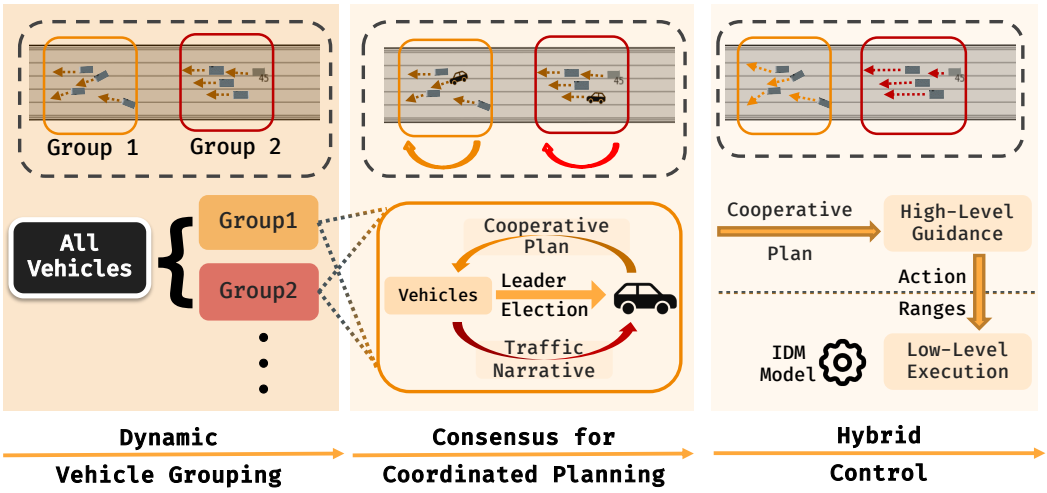

Figure 1: Overview of our LLM-Raft framework, which operates in three stages. Initially, vehicles are dynamically organized into groups based on their proximity. The central stage is consensus for coordinated planning: a leader gathers individual intents, or "traffic narratives", from all members, uses an LLM to reconcile them into a unified cooperative plan, and then ensures group-wide alignment on this plan using a Raft-inspired agreement mechanism. Finally, a hybrid controller executes the agreed-upon plan for each vehicle.

like position and velocity, but cannot operate on the rich, semantic intent—the "why" behind an action—that LLMs can generate. This creates a gap: we have highly intelligent agents, but we lack a coordination method that can harness their full reasoning power.

To bridge this gap, we propose **LLM-Raft**, a novel framework for decentralized coordination among LLM-powered vehicles in urban settings. Inspired by the Raft consensus algorithm [22], LLM-Raft enables vehicles to dynamically form groups and efficiently achieve consensus on high-level "traffic narratives". These narratives are structured, explicitly human-interpretable propositions that clearly describe a vehicle's intent and justification (e.g., "I am yielding to an approaching ambulance for safety"). By mutually agreeing on these semantic plans, groups of vehicles can execute complex maneuvers in a coherent and safe manner without a central authority. A hybrid control system then translates these high-level strategic agreements into real-time actions, as illustrated in Figure 1.

Our work makes the following contributions to the field of urban AI:

- We propose a framework to address the critical challenge of decentralized AV coordination in complex and dynamic urban environments.
- We introduce LLM-Raft, a novel approach that uses semantic consensus on "traffic narratives" to enable more intelligent and human-like coordination among vehicles.
- We empirically demonstrate in realistic urban simulations that our approach significantly improves traffic safety and efficiency, offering a practical path toward more sustainable and intelligent urban mobility.

2 LLM-Raft

To enable effective decentralized coordination, our LLM-Raft framework operates in a three-stage process, as shown in Figure 1. It first dynamically groups nearby vehicles, then establishes a shared plan through semantic consensus, and finally executes the plan using a hybrid controller. This hierarchical design allows us to leverage the advanced reasoning of LLMs for high-level strategy while relying on proven, real-time controllers for safe execution.

2.1 Dynamic Vehicle Grouping

To handle the fluid nature of city traffic, vehicles first organize themselves into small, dynamic groups based on proximity. This allows for localized coordination without overwhelming the communication network. Groups form when vehicles are close enough to cooperate effectively, merge when they overlap, and divide when they move apart. This adaptive structure ensures that coordination is always relevant to the immediate traffic context. The implementation for grouping is detailed in Appendix C.

2.2 Consensus for Coordinated Planning

This stage is the core of LLM-Raft, where vehicles in a group agree on a unified plan. This is achieved through a four-step process inspired by the Raft algorithm [22], with pseudocode in Appendix D:

1. **Individual Plan Proposal.** Each vehicle uses its onboard LLM to generate a "traffic narrative". This is a structured, human-like statement of its intent and reasoning (e.g., "I plan to merge left to overtake a slow vehicle").
2. **Leader Election.** The group elects a temporary leader using a simple heartbeat mechanism. All other "follower" vehicles send their traffic narratives to this leader. This avoids chaotic all-to-all communication and streamlines decision-making.
3. **Plan Reconciliation.** The leader's LLM acts as a group strategist. It aggregates all proposed narratives and synthesizes them into a single, unified cooperative plan that is safe and efficient for the entire group. This plan includes a high-level group narrative and specific action ranges (acceleration, steering) for each vehicle. The prompt is structured in Appendix E.
4. **Unified Plan Agreement.** The leader broadcasts the final cooperative plan to all followers. Followers acknowledge and commit to this plan, ensuring the entire group operates from a single source of truth before any actions are taken.

2.3 Hybrid Control and Execution

Once a plan is agreed upon, it is executed using a hybrid control model. The action ranges from the LLM's cooperative plan serve as high-level strategic guidance. At the low level, a simple and reliable reactive controller, the Intelligent Driver Model (IDM), operates at high frequency. The IDM makes second-by-second driving decisions but is constrained to stay within the strategic action ranges set by the group plan. This ensures that each vehicle's actions are both locally reactive and globally coordinated, as detailed in Appendix F.

2.4 Fault Tolerance

To ensure robustness for real-world urban deployment, LLM-Raft includes simple fault tolerance mechanisms. If a group repeatedly fails to elect a leader or the leader fails to produce a plan within a time limit, the consensus process times out. Vehicles then revert to independent, safe-driving behavior until they can join a new, functional group. This prevents system stalls and ensures predictable behavior during communication failures.

3 Experiment

We designed our experiments to evaluate LLM-Raft 's effectiveness in realistic urban traffic scenarios. Our goal was to answer three key questions: (1) Does our coordinated approach improve traffic efficiency and safety compared to uncoordinated, single-agent methods? (2) Is the semantic consensus mechanism the primary reason for these improvements? (3) How robust is our framework in safety-critical urban emergencies?

3.1 Experiment Setup

We conduct our evaluation in two distinct simulators, MetaUrban [38] and LimSim [37], to test our framework across a range of urban conditions. We vary traffic density from sparse to dense and tested performance in several challenging scenarios, including crowded roads and intersections. A detailed description of the environments, baselines, and evaluation metrics can be found in the AppendixG.

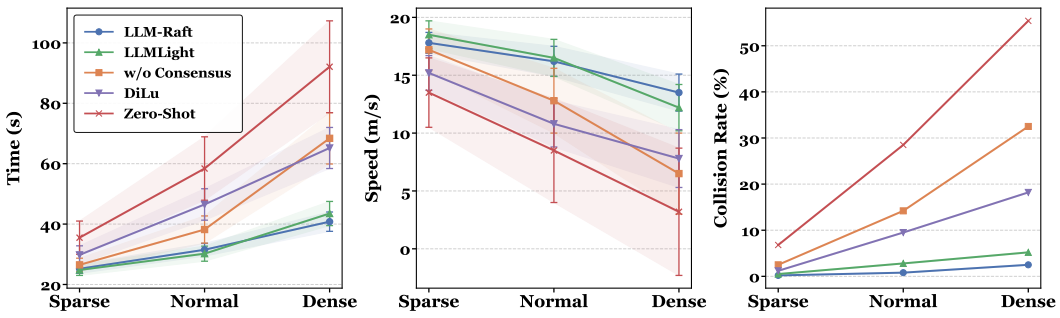

Figure 2: Performance comparison of different methods across varying traffic densities in MetaUrban. The results highlight two key findings: LLM-Raft with its consensus mechanism not only consistently outperforms all baseline methods, but its advantages—particularly the significant reduction in collision rates—become more pronounced as traffic density increases.

3.2 Main Results: Improving Urban Traffic Flow

Our results show that LLM-Raft delivers significant improvements in both traffic efficiency and safety, with the benefits becoming more pronounced as traffic density increases. As shown in Figure 2, while all methods perform reasonably well in sparse traffic, LLM-Raft 's advantage grows in normal and dense conditions, which better represent typical urban congestion.

Specifically, in dense traffic, LLM-Raft reduces task completion time by 20-30% compared to the baselines. More importantly, it achieves a major reduction in collision rates—by 40-50%—demonstrating its value in making urban roads safer. This shows that coordinated planning is critical for navigating complex, crowded environments.

3.3 The Value of Consensus: An Ablation Study

To confirm that these gains come from our core idea, we compared our full framework against the version without consensus. The results in Figure 6 in Appendix H, are clear: enabling semantic consensus provides consistent and significant benefits across all tested scenarios and LLM backbones.

3.4 Robustness in Urban Emergency Scenarios

Effective urban transportation systems must be robust, especially in emergencies. We tested our framework in a scenario where pedestrians suddenly appear on the road, forcing vehicles to coordinate a rapid, collective stop.

As shown in Figure 7 in Appendix H, LLM-Raft significantly outperforms all other methods. By quickly forming a consensus on how to react, our framework enabled a faster group response, achieving a 20-50% faster reaction time. This superior coordination led to a much higher success rate (95.8% vs. 82.7-89.2% for baselines) and maintained better traffic flow during the event. This highlights the practical value of semantic consensus in safety-critical urban situations.

4 Conclusion

In this paper, we addressed the critical challenge of coordinating LLM-powered autonomous vehicles in complex urban environments. We introduced LLM-Raft , a novel framework that enables decentralized coordination through semantic consensus on "traffic narratives". Our experiments demonstrate that by allowing vehicles to agree on high-level, shared plans, our approach significantly improves traffic efficiency and safety. Specifically, LLM-Raft reduced task completion times by up to 20-30% and, most importantly, lowered collision rates by 40-50% in dense traffic compared to uncoordinated methods. This work shows a promising path toward developing more intelligent, collaborative, and resilient multi-agent systems, paving the way for safer and more efficient future urban transportation.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Miguel Castro, Barbara Liskov, et al. Practical byzantine fault tolerance. In *OsDI*, volume 99, pages 173–186, 1999.

[3] Long Chen, Oleg Sinavski, Jan Hünermann, Alice Karnsund, Andrew James Willmott, Danny Birch, Daniel Maund, and Jamie Shotton. Driving with llms: Fusing object-level vector modality for explainable autonomous driving. *arXiv preprint arXiv:2310.01957*, 2023.

[4] Shanzhi Chen, Jinling Hu, Yan Shi, Ying Peng, Jiayi Fang, Rui Zhao, and Li Zhao. Vehicle-to-everything (v2x) services supported by lte-based systems and 5g. *IEEE Communications Standards Magazine*, page 70–76, Jan 2017. doi: 10.1109/mcomstd.2017.1700015. URL `http://dx.doi.org/10.1109/mcomstd.2017.1700015`.

[5] Xiang Cheng, Chen Chen, Wuxiong Zhang, and Yang Yang. 5g-enabled cooperative intelligent vehicular (5genciv) framework: When benz meets marconi. *IEEE Intelligent Systems*, 32(3): 53–59, May 2017. doi: 10.1109/mis.2017.53. URL `http://dx.doi.org/10.1109/mis.2017.53`.

[6] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Receive, reason, and react: Drive as you say with large language models in autonomous vehicles. *arXiv preprint arXiv:2310.08034*, 2023.

[7] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 902–909, 2024.

[8] Thanh-Son Dao, Christopher Michael Clark, and Jan Paul Huissoon. Distributed platoon assignment and lane selection for traffic flow optimization. In *2008 IEEE Intelligent Vehicles Symposium*, Jun 2008. doi: 10.1109/ivs.2008.4621202. URL `http://dx.doi.org/10.1109/ivs.2008.4621202`.

[9] Pedro Fernandes and Urbano Nunes. Platooning with ivc-enabled autonomous vehicles: Strategies to mitigate communication delays, improve safety and traffic flow. *IEEE Transactions on Intelligent Transportation Systems*, 13(1):91–106, Mar 2012. doi: 10.1109/tits.2011.2179936. URL `http://dx.doi.org/10.1109/tits.2011.2179936`.

[10] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[11] Dirk Helbing. Traffic and related self-driven many-particle systems. *Reviews of Modern Physics*, page 1067–1141, Jul 2002. doi: 10.1103/revmodphys.73.1067. URL `http://dx.doi.org/10.1103/revmodphys.73.1067`.

[12] Dongyao Jia and Dong Ngoduy. Enhanced cooperative car-following traffic model with the combination of v2v and v2i communication. *Transportation Research Part B: Methodological*, page 172–191, Aug 2016. doi: 10.1016/j.trb.2016.03.008. URL `http://dx.doi.org/10.1016/j.trb.2016.03.008`.

[13] Dongyao Jia, Kejie Lu, Jianping Wang, Xiang Zhang, and Xuemin Shen. A survey on platoon-based vehicular cyber-physical systems. *IEEE communications surveys & tutorials*, 18(1): 263–284, 2015.

[14] Ali Keysan, Andreas Look, Eitan Kosman, Gonca Gürsun, Jörg Wagner, Yu Yao, and Barbara Rakitsch. Can you text what is happening? integrating pre-trained language encoders into trajectory prediction models for autonomous driving. *arXiv preprint arXiv:2309.05282*, 2023.

[15] Siqi Lai, Zhao Xu, Weijia Zhang, Hao Liu, and Hui Xiong. Llmlight: Large language models as traffic signal control agents, 2024.

[16] Leslie Lamport. The part-time parliament. In *Concurrency: the Works of Leslie Lamport*, pages 277–317. 2019.

[17] Leslie Lamport. Time, clocks, and the ordering of events in a distributed system. In *Concurrency: the Works of Leslie Lamport*, pages 179–196. 2019.

[18] Yongfu Li, Chuancong Tang, Kezhi Li, Xiaozheng He, Srinivas Peeta, and Yibing Wang. Consensus-based cooperative control for multi-platoon under the connected vehicles environment. *IEEE Transactions on Intelligent Transportation Systems*, 20(6):2220–2229, Jun 2019. doi: 10.1109/tits.2018.2865575. URL http://dx.doi.org/10.1109/tits.2018.2865575.

[19] Jiaqi Liu, Peng Hang, Jianqiang Wang, Jian Sun, et al. Mtd-gpt: A multi-task decision-making gpt model for autonomous driving at unsignalized intersections. *arXiv preprint arXiv:2307.16118*, 2023.

[20] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023.

[21] Wenjie Luo, Bin Yang, and Raquel Urtasun. Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. doi: 10.1109/cvpr.2018.00376. URL http://dx.doi.org/10.1109/cvpr.2018.00376.

[22] Diego Ongaro and John Ousterhout. In search of an understandable consensus algorithm. In *2014 USENIX annual technical conference (USENIX ATC 14)*, pages 305–319, 2014.

[23] OpenAI. Gpt-4.1, 2025. URL https://openai.com/index/gpt-4-1. Accessed: 2025-08-01.

[24] Yiyuan Pan, Yunzhe Xu, Zhe Liu, and Hesheng Wang. Planning from imagination: Episodic simulation and episodic memory for vision-and-language navigation, 2024. URL https://arxiv.org/abs/2412.01857.

[25] Maurizio Porfiri, D. Gray Roberson, and Daniel J. Stilwell. Tracking and formation control of multiple autonomous agents: A two-level consensus approach. *Automatica*, 43(8):1318–1328, Aug 2007. doi: 10.1016/j.automatica.2007.01.004. URL http://dx.doi.org/10.1016/j.automatica.2007.01.004.

[26] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[27] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[28] Wei Ren. Consensus based formation control strategies for multi-vehicle systems. In *2006 American Control Conference*, Jan 2006. doi: 10.1109/acc.2006.1657384. URL http://dx.doi.org/10.1109/acc.2006.1657384.

[29] Yao Sun, Lei Zhang, Gang Feng, Bowen Yang, Bin Cao, and Muhammad Ali Imran. Blockchain-enabled wireless internet of things: Performance analysis and optimal communication node deployment. *IEEE Internet of Things Journal*, page 5791–5802, Jun 2019. doi: 10.1109/jiot.2019.2905743. URL http://dx.doi.org/10.1109/jiot.2019.2905743.

[30] Kanata Suzuki and Tetsuya Ogata. Sensorimotor attention and language-based regressions in shared latent variables for integrating robot motion learning and llm. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11872–11878, 2024. doi: 10.1109/IROS58592.2024.10802349.

[31] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[32] Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[33] M. Torrent-Moreno, J. Mittag, P. Santi, and H. Hartenstein. Vehicle-to-vehicle communication: Fair transmit power control for safety-critical information. *IEEE Transactions on Vehicular Technology*, 58(7):3684–3703, Sep 2009. doi: 10.1109/tvt.2009.2017545. URL http://dx.doi.org/10.1109/tvt.2009.2017545.

[34] Martin Treiber, Ansgar Hennecke, and Dirk Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical review E*, 62(2):1805, 2000.

[35] Ziran Wang, Guoyuan Wu, and Matthew Barth. A review on cooperative adaptive cruise control (cacc) systems: Architectures, controls, and applications. *Cornell University - arXiv,Cornell University - arXiv*, Sep 2018.

[36] Licheng Wen, Daocheng Fu, Xin Li, Xinyu Cai, Tao Ma, Pinlong Cai, Min Dou, Botian Shi, Liang He, and Yu Qiao. Dilu: A knowledge-driven approach to autonomous driving with large language models. *arXiv preprint arXiv:2309.16292*, 2023.

[37] Licheng Wenl, Daocheng Fu, Song Mao, Pinlong Cai, Min Dou, Yikang Li, and Yu Qiao. Limsim: A long-term interactive multi-scenario traffic simulator. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1255–1262. IEEE, 2023.

[38] Wayne Wu, Honglin He, Jack He, Yiran Wang, Chenda Duan, Zhizheng Liu, Quanyi Li, and Bolei Zhou. Metaurban: An embodied ai simulation platform for urban micromobility. *ICLR*, 2025.

[39] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*, 2024.

[40] Yinlong Zhang, Wei Liang, Sichao Zhang, Xudong Yuan, Xiaofang Xia, Jindong Tan, and Zhibo Pang. High-precision calibration of camera and imu on manipulator for bio-inspired robotic system. *Journal of Bionic Engineering*, 19(2):299–313, 2022. ISSN 2543-2141. doi: 10.1007/s42235-022-00163-7. URL https://doi.org/10.1007/s42235-022-00163-7.

[41] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023.

A Limitations and Future Work

While our results demonstrate the promise of the LLM-Raft framework, we acknowledge several limitations that present important avenues for future research.

- **Scalability of the Leader-Based Model:** In our current implementation, a single leader is responsible for reconciling plans for its group. While effective for small to medium-sized groups, this approach could become a communication and computational bottleneck in extremely dense urban scenarios with very large groups. Future work could explore more scalable consensus structures, such as hierarchical leader elections or fully decentralized protocols where the reconciliation workload is distributed among all members.
- **Reliance on LLM Performance:** The effectiveness of our framework is inherently tied to the performance of the underlying Large Language Model, including its reasoning accuracy and inference latency. For real-world deployment on vehicles with limited computational resources, research into model compression, distillation, and quantization techniques tailored for this task will be essential to ensure low-latency, high-fidelity performance.
- **Simulation-to-Reality Gap:** Our current work validates the framework in simulation using text-based "traffic narratives". Real-world urban environments are far more complex and uncertain. Bridging the sim-to-real gap will require extending these narratives to a richer, multi-modal format that can incorporate real-time sensor data from cameras, LiDAR, and radar. This would enable more robust coordination in the face of unpredictable events and noisy sensor inputs.
- **Communication Security:** The framework assumes a reliable and secure V2X communication channel. In practice, vehicular networks are vulnerable to various security threats, including message spoofing, replay attacks, and denial-of-service attacks. Future iterations of the framework should incorporate security protocols to verify the identity of participating vehicles and ensure the integrity of the exchanged traffic narratives.

B Related Work

B.1 Consensus in Vehicular Networks

Consensus algorithms are essential for distributed systems to reach agreement in spite of failures or unreliable communication. Foundational protocols like Paxos [16] and Practical Byzantine Fault Tolerance (PBFT) [2] provide robust mechanisms for ensuring data consistency and are widely used in fault-tolerant computing [17].

In the context of autonomous systems, these algorithms have been adapted for various vehicular applications, such as Cooperative Adaptive Cruise Control (CACC) and vehicle platooning [13, 35]. This line of research focuses on exchanging low-level parameters (e.g., speed, spacing) to maintain stable formations and improve traffic flow [8, 9, 11, 12, 28]. Advanced strategies like multi-platoon coordination enhance road efficiency but are limited in handling complex maneuvers [18, 25]. While effective for synchronizing vehicle movements, these methods operate purely on kinematic data. They are not designed to understand or coordinate based on the underlying intent or context of vehicle actions, thus failing to address the semantic gap.

B.2 LLMs for Autonomous Driving

Large Language Models (LLMs) have shown significant promise in robotics and autonomous systems [40, 30] due to their advanced reasoning capabilities [1, 27]. In autonomous driving, research has largely focused on using LLMs to improve the decision-making of a single, ego-centric vehicle [1, 24, 26, 27, 21, 41]. These approaches can be broadly categorized into two types: fine-tuning and prompt engineering.

Fine-tuning approaches adapt pre-trained LLMs to driving-specific tasks. For example, DriveGPT4 is a multi-modal model fine-tuned for end-to-end driving tasks [39], while other works convert decision-making into a sequence modeling problem [3, 19, 20]. These methods demonstrate LLMs can handle complex scenarios, but they are designed for a single agent and do not address multi-vehicle coordination.

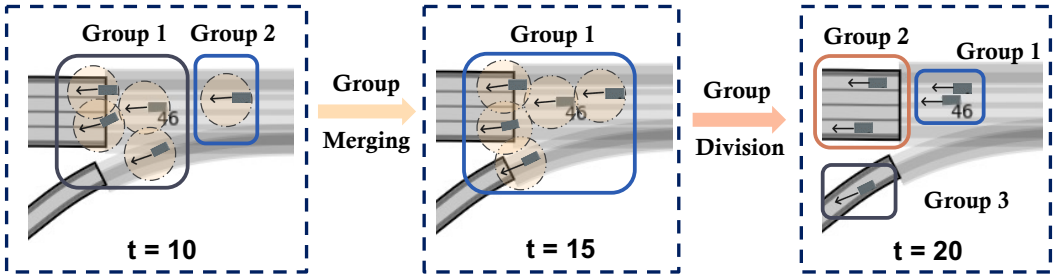

Figure 3: The dynamic vehicle grouping process. Groups are formed based on vehicle proximity, allowing for flexible merging and division as traffic conditions evolve. The figure illustrates two operations: (a) two separate groups converging and merging into a larger one (t=10 to t=15) , and (b) an isolated vehicle subsequently detaching to form a new, independent group (t=20).

Prompt engineering approaches leverage the zero-shot reasoning abilities of LLMs. Frameworks like DiLu use carefully designed prompts to enable an LLM to act as a "driver agent" for a single vehicle [36]. Other methods focus on creating a chain-of-thought process for the LLM to reason about its actions [6, 7, 14]. While powerful, these works treat each vehicle as an isolated agent making individual decisions based on its own observations and reasoning.

On one hand, research on consensus algorithms in vehicular networks creates robust protocols for data synchronization but overlooks the underlying semantic meaning of the data. On the other hand, research on LLMs for autonomous driving focuses on enhancing individual vehicle decision-making but does not explore how to use LLMs to achieve coordinated group behavior.

C Detailed Implementation of Dynamic Vehicle Grouping

This section provides a detailed description of the dynamic vehicle grouping mechanism used in the LLM-Raft framework, as shown in Figure 3. The goal of this mechanism is to create context-aware, localized coordination groups that can adapt to the constantly changing conditions of urban traffic.

C.1 Principle of Proximity-Based Grouping

In our framework, vehicles form coordination groups based on their real-time spatial proximity. A group is defined as a set of vehicles where every member is close enough to at least one other member to engage in meaningful coordination.

Formally, given a set of vehicles $V = \{v_1, v_2, ..., v_n\}$ with positions $\mathbf{p}_i(t) \in \mathbb{R}^2$ at time t, a group $G_k \subseteq V$ is formed if for any vehicle $v_i \in G_k$, there exists at least one other vehicle $v_j \in G_k$ (where $i \neq j$) such that the distance between them is less than or equal to a predefined threshold, D_{thresh}.

In our implementation, we set $D_{\text{thresh}} = 30$ meters. This value was empirically chosen to balance reliable V2X (Vehicle-to-Everything) communication range with the practical need for coordination in typical urban driving scenarios.

C.2 Group Merging and Division for Dynamic Adaptation

To adapt to fluid traffic, groups are not static. They can merge when they get close or divide as vehicles move apart.

- **Group Merging:** Two distinct groups, G_i and G_j, merge into a single, larger group when the minimum distance between any vehicle in G_i and any vehicle in G_j drops below the merging threshold, D_{merging}. We set $D_{\text{merging}} = 20$ meters.
- **Group Division:** A vehicle v_i separates from its current group G_k if its distance to every other member in G_k exceeds the division threshold, D_{division}. The departing vehicle may then form a new group on its own or join another nearby group. We set $D_{\text{division}} = 40$ meters.

Algorithm 1 Consensus for Vehicle Coordination

Require: data $= \{\mathcal{V}, a, \omega, E\}$
Ensure: Synchronization
$\forall v \in \mathcal{V}, v \leftarrow$ follower
while $\neg\exists$ leader **do**
  $\mathcal{V} \xrightarrow{\text{elect}} v_{\text{leader}}$
end while
while leader_active **do**
  $\forall v \in \mathcal{V}, v \xrightarrow{\text{heartbeat}} v_{\text{leader}}$
  if $\neg$heartbeat **then**
    re-elect(v_{leader})
  end if
  $\mathcal{V}_{\text{follower}} \xrightarrow{\text{intent, data}} v_{\text{leader}}$
  $v_{\text{leader}} \xrightarrow{\text{process}} \text{LLM}(\text{data})$
  $\text{LLM} \xrightarrow{\text{cooperative plan}} \mathcal{V}_{\text{follower}}$
end while

C.3 Hysteresis for Stability

You may notice that the merging threshold (20m) is smaller than the division threshold (40m). This is intentional and creates a hysteresis effect. This design prevents unstable situations where vehicles at the boundary of the communication range might rapidly join and leave a group. By requiring vehicles to be significantly further apart to divide than to merge, we ensure that group compositions remain stable for longer, leading to more consistent and reliable coordination.

D Pseudocode of Consensus in LLM-Raft

E Prompt of LLMs for Plan Reconciliation

```
Example Prompt (a) Role of LLM You are an expert driving strategist for a group of communicating
ve-hicles. Your primary goal is to ensure safety and traffic efficiency. (b) Input Context: our own
ve-hicle's state and proposed plan is [Self Narrative: Intention, Velocity, Acceleration, Position
coordinates, Angular velocity, etc.]
The other vehicles in your group have proposed the following plans: [List of narratives from followers:
Surrounding vehicle speeds & Coordinates, Surrounding count, etc.]
Information about the road you are driving on: [Environmental Information: Intersection distance, Road
complexity, Road length, Traffic light timing, etc.] (c) Task description Review all proposed plans.
Re-solve any conflicts and generate a single, unified cooperative plan for the entire group. (d)
Output format Provide the full cooperative plan in a structured JSON format.
Provide the acceleration and angular velocity that you think is suitable for the current scenario: [Acc
Range] & [Steering Range]
```

F Detailed Implementation of Hybrid Control and Execution

To address the numerical instability of pure LLM-based control and the lack of adaptability in purely rule-based methods, our framework uses a hybrid approach that combines the strengths of both. The process unfolds in three stages.

F.1 High-Level Strategic Guidance from the LLM

First, the LLM acts as a high-level strategic planner. After the consensus process, the leader LLM generates a unified cooperative plan. For each vehicle, this plan specifies a set of permissible action ranges, which serve as strategic guidelines. A typical output for a single vehicle is a JSON object defining these ranges:

Figure 4: MetaUrban environment visualization across different traffic densities. Top row shows bird's-eye view perspectives, bottom row shows third-person view perspectives. From left to right: sparse traffic (5-8 vehicles), normal traffic (12-18 vehicles), and dense traffic (25-35 vehicles).

```
{
  "vehicle_id": "Car_A",
  "acceleration_range": [-5.0, 10.0],
  "steering_range": [-40.0, 40.0]
}
```

Here:

- **acceleration_range:** Defines the minimum and maximum permissible longitudinal acceleration in m/s^2. A negative value indicates braking.
- **steering_range:** Defines the minimum and maximum permissible steering angle in radians.

These ranges are not specific commands but rather a strategic "envelope" or "guideline". They represent the set of actions that are considered safe and consistent with the group's overall cooperative plan for the next time step.

F.2 Low-Level Control Refinement with IDM

Next, the Intelligent Driver Model (IDM) [34] converts these strategic ranges into precise, executable control signals.

For longitudinal control, the IDM calculates the vehicle's acceleration $\dot{v}$ using the following equations:

$$\begin{aligned} \dot{v} &= a\left[1-\left(\frac{v}{v_0}\right)^{\delta}-\left(\frac{s^*(v,\Delta v)}{s}\right)^2\right], \\ s^*(v,\Delta v) &= s_0+\max\left(0, vT+\frac{v\Delta v}{2\sqrt{ab}}\right), \end{aligned} \tag{1}$$

where v is the current speed, v_0 is the target speed, s is the actual distance to the lead vehicle, and s_0 is the minimum safe following distance. The parameter δ is the acceleration index.

Crucially, the LLM's output directly constrains the IDM's behavior. The maximum acceleration a and the comfortable braking deceleration b used in the IDM calculation are taken from the

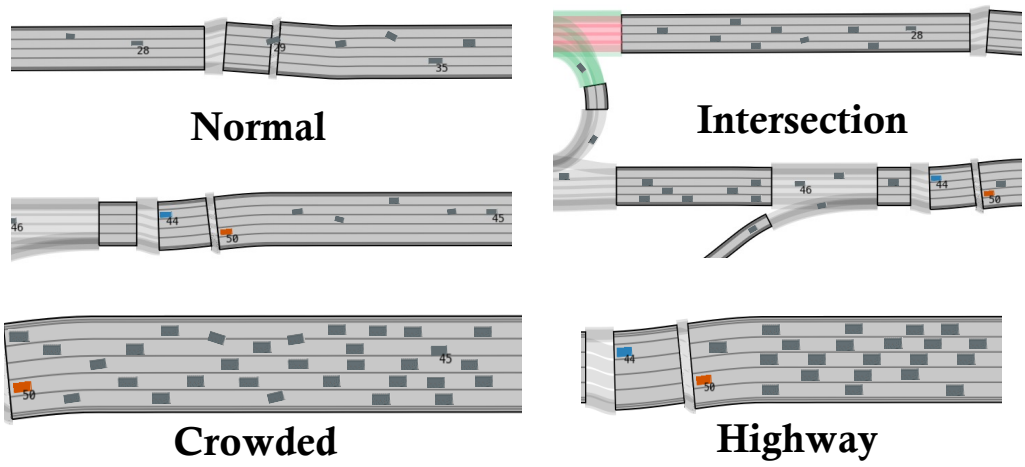

Figure 5: LimSim environment scenarios with increasing coordination complexity. Four driving scenarios are designed to test multi-vehicle coordination: Normal Road (10-15 vehicles), Crowded Road (30-50 vehicles), Highway (20-30 vehicles), and Intersection (20-30 vehicles), each presenting distinct coordination challenges for autonomous vehicles requiring different levels of inter-vehicle communication.

'acceleration_range' provided by the LLM. This ensures that the vehicle's real-time, reactive behavior always adheres to the high-level group strategy.

For lateral control, the final steering angle is determined by first ensuring the steering direction is within the LLM's permissible 'steering_range' and then averaging the allowable range of steering angles to produce a smooth maneuver.

This hybrid approach allows LLM-Raft to benefit from the contextual reasoning of the LLM for multi-agent planning while relying on the proven stability of the IDM for safe and smooth vehicle operation.

G Detailed Experimental Setup

This section provides a comprehensive overview of the simulation environments, baseline methods, and evaluation metrics used to validate the LLM-Raft framework.

G.1 Simulation Environments and Scenarios

To ensure a thorough and robust evaluation, we used two complementary simulation platforms with distinct characteristics.

- **MetaUrban (Figure 4):** A lightweight yet powerful simulator known for its ability to generate complex, dense urban traffic and support multi-modal inputs [38]. Its high-fidelity environment makes it ideal for head-to-head performance comparisons and for testing robustness in safety-critical situations.
  - **Traffic Density Scenarios:** We tested three levels of traffic density: Sparse (5-8 vehicles), Normal (12-18 vehicles), and Dense (25-35 vehicles).
  - **Emergency Scenario:** A specialized scenario in the "Normal" density setting where 3-5 pedestrians suddenly cross the road, designed to test the system's crisis management capabilities.

- **LimSim (Figure 5):** A lightweight simulator focused purely on traffic planning [37]. Its flexibility allows for easy customization of diverse road layouts, making it perfect for controlled ablation studies that isolate the performance of the consensus mechanism.

– **Coordination Complexity Scenarios:** We designed four scenarios with increasing coordination challenges: Normal Road (10-15 vehicles), Crowded Road (30-50 vehicles), Highway (20-30 vehicles), and Intersection (20-30 vehicles).

In all scenarios, a portion of the vehicles (30-70%) are designated with specific destination goals, while the rest act as background traffic controlled by default rule-based policies.

G.2 Baseline Methods

To demonstrate the value of our coordination framework, we compare LLM-Raft against representative LLM-based driving methods that rely on text-based inputs. In all baseline experiments, each vehicle makes decisions independently based on its own observations, without any inter-vehicle communication or coordination.

- **LLMLight [15] and DiLu [36]:** These represent the current state-of-the-art in LLM-powered autonomous agents. They leverage advanced reasoning, memory modules, and prompt engineering to navigate complex traffic environments. Specifically, LLMLight demonstrates the capability of LLMs in traffic flow optimization, while DiLu utilizes a memory bank to learn from past driving experiences.
- **Zero-shot LLM:** A fundamental baseline where the LLM makes driving decisions directly based on the current observation prompt, without access to any memory module, few-shot examples, or coordination mechanisms. This serves as a lower bound to demonstrate the difficulty of the task without specialized framework support.
- **LLM-Raft (w/o Consensus):** This is a critical ablation of our own framework where we disable the consensus mechanism. Each vehicle still uses the LLM-Raft architecture for planning but does so individually without communicating or agreeing on a group plan. This baseline allows us to isolate and quantify the direct performance contribution of the semantic consensus process.

G.3 LLMs Used in the Experiments

To test the generalizability of our framework, we evaluate it across four large language models with varying capabilities: GPT-4.1 [23], Gemini-2.5-Pro [31], DeepSeek-R1 [10], and Qwen3-235B-a22B [32].

For the main results in the more complex MetaUrban environment, we use GPT-4.1 as the primary LLM for all methods. For the ablation studies in LimSim, we ran experiments with all four models to validate that the benefits of the consensus mechanism are independent of the specific LLM used.

G.4 Evaluation Metrics and Criteria

We use a set of standard metrics to assess performance from the perspectives of efficiency, safety, and overall task success. All results are averaged over 30 independent trials for each configuration.

- **Task Completion Time (seconds):** The time taken for all designated vehicles to reach their destinations. Lower is better.
- **Average Speed (m/s):** The average speed of all designated vehicles throughout the trial. Higher is better, as it indicates smoother traffic flow.
- **Collision Rate (%):** The percentage of trials in which any inter-vehicle collision occurred. This is our primary metric for safety. Lower is better.
- **Success Rate (%):** The percentage of trials where all designated vehicles successfully reached their destinations without any collisions or rule violations within the 100-second time limit. Higher is better.

A trial is considered a failure if any of the following occurs:

- Any designated vehicle fails to reach its destination within the time limit.
- Any collision occurs between vehicles, or between a vehicle and a pedestrian or obstacle.
- A major traffic rule is violated (e.g., driving off the road, running a red light).

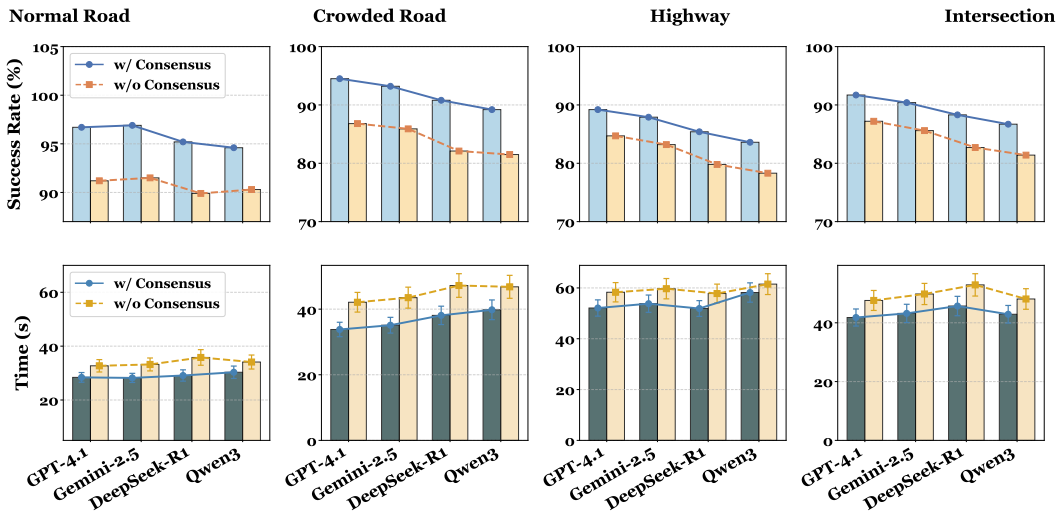

Figure 6: Performance comparison of our framework with and without the consensus mechanism across four distinct scenarios in LimSim. The results demonstrate that the consensus mechanism provides a consistent performance benefit that is both independent of the LLM's capability and becomes more pronounced as the coordination complexity of the scenario increases.

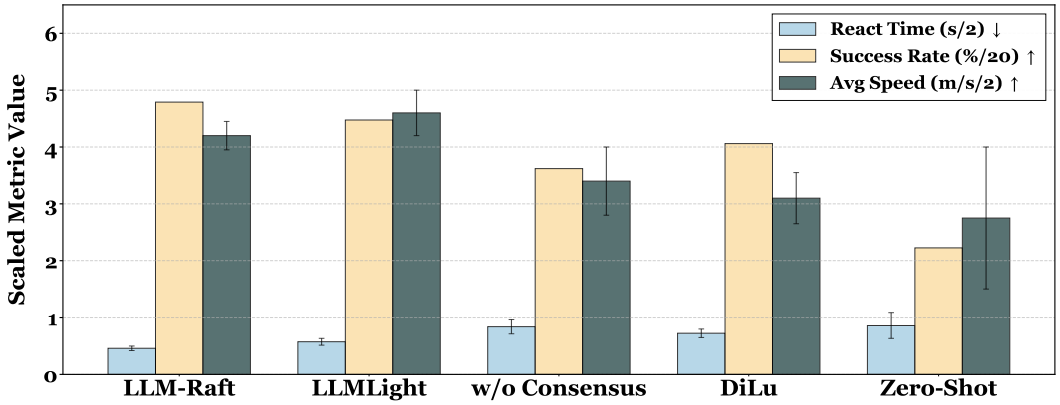

Figure 7: Performance in MetaUrban's emergency scenarios. LLM-Raft significantly outperforms all baseline methods by enabling a rapid, coordinated group response.

H More Results of Our Experiment

We place additional experiment results here.