A Learning-Based Integrated Framework of Motion Prediction and Planning for Connected and Automated Vehicles: Towards Interaction, Multi-Modality, and Relational Reasoning

by

Keshu Wu

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Civil and Environmental Engineering)

at the

UNIVERSITY OF WISCONSIN–MADISON

2024

Date of final oral examination: 04/15/2024

The dissertation is approved by the following members of the Final Oral Committee:
Dr. Bin Ran, Professor, Civil and Environmental Engineering, UW-Madison
Dr. David Noyce, Professor, Civil and Environmental Engineering, UW-Madison
Dr. Sue Ahn, Professor, Civil and Environmental Engineering, UW-Madison
Dr. Xin Wang, Assistant Professor, Industrial and Systems Engineering, UW-Madison
Dr. Yang Zhou, Assistant Professor, Civil and Environmental Engineering, Texas A&M

To my entire family,

thank you for inspiring me to lead a life of wisdom and love.

Acknowledgements

Looking back on my five-year Ph.D. journey in the Department of Civil and Environmental Engineering at the University of Wisconsin-Madison, I am filled with immense gratitude. I received strong support from a stimulating environment that encouraged new ideas and personal growth. The tough academic challenges I faced were balanced by the invaluable friendships and guidance I found along the way. These experiences, along with countless hours of hard work, have played a key role in shaping me as a scholar and researcher, deeply impacting my academic and professional goals.

First and foremost, I wish to extend my deepest appreciation to my advisor, Prof. Bin Ran. His exceptional mentorship, unwavering support, and persistent encouragement have been the cornerstone of my Ph.D. journey. His profound insights and expert feedback have significantly shaped the course of this dissertation. Prof. Ran's ability to inspire and challenge me to reach new heights has been instrumental in my academic and personal growth. His dedication to excellence, passion for work, and critical thinking have set a standard that I will strive to uphold throughout my career.

I am also profoundly grateful to my committee members, Prof. Bin Ran, Prof. David Noyce, Prof. Sue Ahn, Prof. Xin Wang, and Prof. Yang Zhou. Their invaluable time, insightful suggestions, and constructive criticisms have been crucial in refining and enhancing this work.

A special note of thanks goes to my co-authors and colleagues, Prof. Yang Zhou, Prof. Xin Wang, Dr. Steven Parker, Prof. Xiaopeng Li, Dr. Yang Cheng, Dr. Pei Li, and Dr. Haotian Shi. Your guidance and collaboration have been pivotal to the success of this research. Prof. Yang Zhou and Dr. Haotian Shi have been constant sources of wisdom and support. Their innovative ideas and practical advice have significantly contributed to the depth and breadth of this dissertation. Dr. Steven Parker, Dr. Yang Cheng, and Dr. Pei Li have provided invaluable guidance on engineering projects at the Traffic Operations and Safety Laboratory, equipping me for a professional career.

I sincerely appreciate the members of UW Madison Transportation for their unwavering support. Special thanks to Dr. Haotian Shi and Sicheng Fu for their incredible friendship and support. Your mentorship and encouragement were vital during the toughest phases of this journey. To my colleagues in the CAVH research group: Dr. Yifan Yao, Jingwen Zhu, Rei Tamaru, Junwei You, Rui Gan, Kexin Tian, Sixu Li, Junyi Ma, Weizhe Tang, Hanif Nazaroedin, Ran Yi, and Han Cao, your camaraderie and collaboration have greatly enriched my research experience. Special recognition goes to the senior fellows: Dr. Xiaotian Li, Dr. Tianyi Chen, Dr. Shen Li, Dr. Shuoxuan Dong, Dr. Kunsong Shi, and Dr. Yuan Zheng, whose guidance has been invaluable. To my fellow researchers: Xinzhi Zhong, Cesar Andriola, Jiwan Jiang, Zilin Huang, Dr. Chunyu Liu, Yifan Tan, Peipei Mao, Hanchu Li, Ben Zhai, Jiannan Mao, Yuting Chen, Xiaoying Yi, and Yunran Di, your contributions and support have been indispensable throughout this journey.

To my friends Yichen Mo, Zihao Zhang, Yize Liu, and Shaocheng Yang, thank you for the joyous moments and support during the challenging times of Covid. Additionally, my gratitude extends to Haolin Wang, Yekai Lu, Deyang Yin, Ken Chen, and Haozheng Wang for their steadfast friendship and encouragement.

My heartfelt thanks go to my girlfriend, Jingwen Zhu, for her unwavering support and love. Her belief in my abilities has been a constant source of strength and motivation. Her understanding and encouragement have instilled in me the confidence to overcome obstacles and strive for excellence. The joy and comfort she brings into my life have made this journey fulfilling, and I am profoundly grateful for her unwavering faith in me.

A special thanks to my cat Sandy, whose companionship during countless late-night coding sessions has provided comfort and solace.

Lastly, to my dad Xiaodong Wu, my mom Xiaoli He, and my grandparents, your endless love, patience, and unwavering support have been the foundation of my strength. Your belief in me and constant encouragement have helped me navigate every challenge. You have instilled in me the values of perseverance, hard work, and integrity, shaping the person I am today. This achievement is a testament to your sacrifices and unwavering faith in my potential. Your support has been my guiding light, providing me with the confidence and determination to pursue my dreams. I am grateful for all of this.

Contents

List of Tables

List of Figures

Abstract

Predicting vehicle trajectories and ensuring safe and efficient trajectory planning are critical for the operational efficiency and safety of automated vehicles, especially on congested multi-lane highways. In these dynamic environments, a vehicle's movement is influenced by its historical behaviors and interactions with surrounding vehicles. These complex interactions result from unpredictable motion patterns, leading to diverse modalities of driving behaviors that necessitate thorough investigation. Additionally, in multi-agent systems, dynamic interactions among agents often display cooperative and competitive behaviors. Such group-wise interactions, though common, are rarely modeled. Traditional methods, while effective in capturing pair-wise interactions, fail to represent the collective influence of groups on each other's behaviors in real-world traffic scenarios. Therefore, modeling the group-wise interactions of multi-modal driving behaviors among multiple agents is essential. In dense traffic conditions, vehicles frequently change lanes, accelerate, decelerate, and engage in complex interactions with other agents. These interactions often involve multiple possible longitudinal and lateral behaviors of various entities influencing each other simultaneously, which cannot be fully captured by considering only pair-wise relationships. Furthermore, the stochastic nature of human behavior adds complexity, requiring models that handle the uncertainty and variability in agent behaviors for safe and efficient driving. Thus, a critical challenge lies in representing and reasoning about the diverse interactions among agents and their multiple possible behaviors to achieve socially inspired automated driving.

This dissertation introduces the **G**raph-based **I**nteraction-aware **M**ulti-modal **T**rajectory **P**rediction (`GIMTP`) framework, designed to probabilistically predict future vehicle trajectories by effectively capturing these interactions. Within this framework, vehicle motions are conceptualized as nodes in a time-varying graph, and traffic interactions are represented by a dynamic adjacency matrix. To comprehensively capture both spatial and temporal dependencies embedded in this dynamic adjacency matrix, the methodology employs the Diffusion Graph Convolutional Network (DGCN), providing a graph embedding of both historical and future states. Additionally, a driving intention-specific feature fusion is implemented, enabling the adaptive integration of historical and future embeddings for enhanced intention recognition and trajectory prediction. This model offers two-dimensional predictions for each mode of longitudinal and lateral driving behaviors and provides probabilistic future paths with corresponding probabilities, addressing the challenges of complex vehicle interactions and multi-modality of driving behaviors. To further facilitate interaction-aware multi-modal motion prediction for multi-agent systems, `GIMTP` is enhanced to **G**raph-based **I**nteraction-aware **R**eliable **A**nticipative **F**easible **F**uture **E**stimator (`GIRAFFE`), which offers multi-modal

predictions by considering the behaviors of multiple vehicles.

Building upon the robust GIRAFFE framework, this dissertation further integrates the **R**elational **H**ypergraph **I**nteraction-informed **N**eural m**O**tion generator and planner (RHINO), a proposed model for motion planning that revolutionizes interaction modeling and relational reasoning for trajectory prediction and planning with multiscale hypergraph representations. RHINO distinguishes itself by surpassing previous methods that primarily consider pair-wise interactions with limited relational insight. It introduces a multiscale hypergraph neural network designed to capture intricate dynamics involving both pair-wise and group-wise interactions across multiple scales. RHINO's multiscale hypergraph is engineered to be trainable, enabling the system to discern more complex interaction patterns within traffic, such as varying group sizes and the nuances of collective behaviors. For interaction representation learning, RHINO adopts a three-element format that facilitates end-to-end learning. This innovative approach allows for explicit reasoning of relational factors, including interaction strength and category, which are crucial for accurate and socially aware motion planning. Furthermore, RHINO is integrated into both a Conditional Variational Autoencoder (CVAE)-based prediction system and enhanced state-of-the-art prediction frameworks to yield socially plausible trajectories grounded in relational reasoning. The efficacy of RHINO in understanding group behavior and discerning interaction dynamics is substantiated through synthetic physics simulations, reflecting its capability to capture group behaviors and reason about the strength and category of interactions. The effectiveness of this motion planning system is validated through extensive experiments on two real-world trajectory prediction datasets. This integrated framework of motion prediction and planning, adopting the GIRAFFE framework and RHINO framework, positions it as a powerful tool in advancing the safety and efficiency of automated vehicle operations, especially in the complex and unpredictable environment of multi-lane highways.

Chapter 1

Introduction

1.1 Background

The development and integration of connected and automated vehicles (CAVs) hold considerable promise for advancing transportation systems in terms of mobility, efficiency, and safety [1, 2, 3]. Vehicle trajectory prediction and planning are critical components of intelligent transportation systems, essential for enhancing traffic safety, reducing congestion, and promoting sustainable transportation [4, 5]. Efficient vehicle operation necessitates a comprehensive understanding of the driving environment and the generation of precise motion predictions for surrounding objects [6, 7]. This understanding is crucial for intelligent decision-making, trajectory planning, and control during CAV operation, ultimately contributing to the development of an intelligent and dependable transportation network [5, 8, 9, 10].

The trajectory of a vehicle is influenced by a variety of factors present in the driving environment. It's clear that a vehicle's movement is determined not only by its past trajectory but also by the movement and actions of surrounding vehicles, especially in congested traffic situations [11, 12]. Navigating through such complex traffic environments with numerous non-cooperative vehicles introduces a significant level of uncertainty due to many hidden variables [13, 14]. Predicting a target vehicle's future trajectory has become increasingly important and is crucial for improving traffic safety [15, 16]. The intricacies of automated driving and the dynamic nature of dense traffic environments present significant challenges for current trajectory prediction research. These challenges are particularly pronounced in three areas: (i) the interactions between vehicles [17, 18], (ii) the multi-modality of driving behaviors [15, 19], and (iii) the underlying relational interactions between vehicles and their behaviors [20, 21, 22].

1.1.1 Interaction between Vehicles

Predicting the future trajectory of a target vehicle is a complex challenge, especially in dynamic and dense driving environments. The movement of a target vehicle is influenced not only by its own historical path but also by the behaviors and motions of surrounding vehicles. These interactions create a highly interdependent system where the motion of one vehicle can significantly impact the trajectories of others [23]. Consequently, it is essential to analyze how neighboring vehicles affect the target vehicle's motion. This requires the ability to infer potential interaction patterns from raw trajectory data, which is a significant challenge. The difficulty lies in accurately capturing and modeling the diverse and often unpredictable ways in

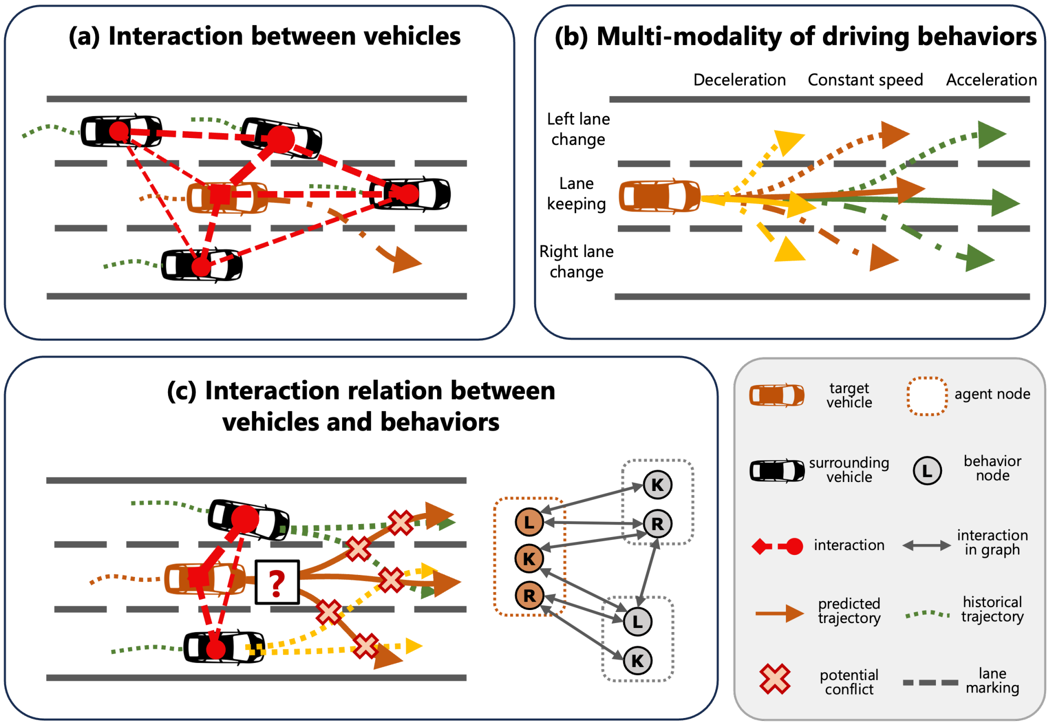


Figure 1.1: Vehicle interaction and behavior multimodality.

which vehicles interact with each other on the road [17].

1.1.2 Multi-modality of Driving Behaviors

The inherent unpredictability and uncertainty of real-world driving scenarios present a major obstacle in accurately predicting a single future trajectory. Various unknown factors, such as individual driver characteristics, and their physical and psychological conditions, lead to different behaviors and reactions among drivers in identical driving situations. For instance, two drivers might respond differently to the same traffic signal, one might slow down gradually, while the other might brake suddenly. Therefore, it is crucial to model the multi-modality of driving intentions. This approach involves generating multiple potential trajectories that represent the range of possible future actions a driver might take. By considering these multiple possible outcomes, prediction models can better account for the variability and uncertainty inherent in driving behaviors, leading to more robust and reliable predictions.

1.1.3 Interaction Relational Reasoning between Vehicles

Predicting the trajectories of multiple agents is crucial in various practical applications, including autonomous driving, mobile robot navigation, and other domains where groups of entities interact, leading to complex behavior patterns at both the individual and system levels. Typically, we have access only to the trajectories of individual entities without any insight into the underlying interaction patterns, and each agent can exhibit multiple possible behavior modalities. This makes modeling these dynamics and predicting future behaviors particularly challenging. Additionally, it is important to study behaviors at the group level, beyond just pair-wise interactions. In multi-agent systems, the dynamic interactions among agents often display cooperative and competitive behaviors [11, 12, 24]. Although these group-level interactions are common, they are rarely modeled. Traditional graph-based methods, while effective at capturing pair-wise interactions, are inadequate in representing the collective influence of groups of entities on each other's behaviors in real-world traffic situations.

As we delve deeper into the intricacies of autonomous driving systems, a prominent challenge that emerges is Interaction Relational Reasoning [20, 21]. The complexity of autonomous driving does not solely lie in the sensor technology or the computational algorithms but equally in the subtleties of interaction between vehicles that share the road.

Autonomous driving systems require a profound understanding of implicit agent interactions. This understanding is crucial because it allows for the anticipation and interpretation of other vehicles' actions without needing a priori knowledge of their underlying intentions. In dense traffic scenarios, vehicles frequently change lanes, accelerate or decelerate, and engage in complex interactions with other agents. These interactions often involve multiple possible longitudinal and lateral behaviors of multiple entities influencing each other simultaneously, which cannot be fully captured by considering only pair-wise relationships. Additionally, the stochastic nature of human behavior adds further complexity, necessitating models that can handle the uncertainty and variability in agent behaviors, which is essential for safe and efficient driving. Therefore, to untangle social influence and achieve socially inspired automated driving, a critical challenge lies in representing and reasoning about the diverse interactions among agents and their multiple possible behaviors [20, 25, 26].

The work by [20] underscores the importance of relational reasoning as a vehicle navigates environments teeming with dynamic and unpredictable elements. Here, the underlying relations that facilitate this silent communication include implicit rules, dependencies, or associations that may exist between different agents or among the different modalities of behaviors of these agents. These relations are "implicit" in that they are neither directly observable nor explicitly communicated.

Instead, they form the hidden tapestry of road interaction that an autonomous vehicle must interpret. This ability to decipher and adapt to these silent rules enables an autonomous vehicle to understand how one agent's intention or behavior could influence another's, thereby ensuring harmonious coexistence on the road.

The explainability in autonomous systems highlights the significance of being able to elucidate the behaviors and intentions of autonomous vehicles [27]. This quest for clarity is pivotal for validating the decisions made by such systems and essential for strategic planning amid the uncertainty of multi-agent interactions. Explainability thus becomes a bedrock for achieving a level of decision-making and planning that mirrors human-like intuition and judgment.

Safety and efficiency, the cornerstones of AV and CAV operations, are directly impacted by the capability of these vehicles to reason relationally. It ensures that the vehicles operate not in isolation but in a state of continuous and fluid dialogue with the surrounding traffic. This dialogue does not occur through explicit communication but rather through a sophisticated understanding of traffic flow, other vehicles' movements, and the silent choreography of the road.

In summary, the relational reasoning and implicit understanding of the complex web of interactions between vehicles play a pivotal role in the present and future of autonomous driving. Advancements in this area will continue to push the boundaries of what is possible, guiding us toward a future where vehicles are not merely machines on the road but active, responsive participants in a larger vehicular society.

1.1.4 The "Social" Nature of Automated Driving

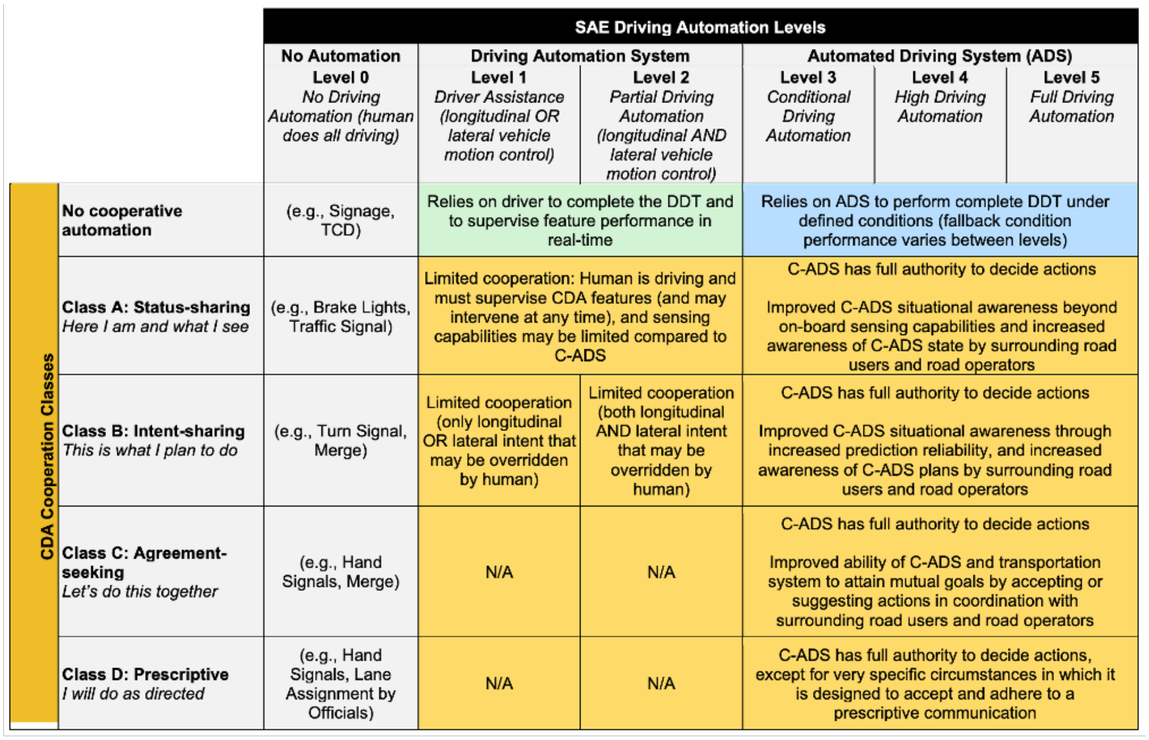
The SAE J3216_202107 standard "Taxonomy and Definitions for Terms Related to Cooperative Driving Automation for On-Road Motor Vehicles" [28] provides a taxonomy and definitions for terms related to Cooperative Driving Automation (CDA) for on-road motor vehicles, marking a significant step toward the realization of fully automated and cooperative driving systems. The classification of CDA is structured into four classes:

1. Status-sharing: At the fundamental level, status-sharing encompasses the communication of a vehicle's state to traffic participants and the environment. This class forms the basis for all higher levels of CDA.

2. Intent-sharing: Intent-sharing involves conveying the intended actions of the Cooperative-ADS (C-ADS) to other traffic participants and understanding their intentions in return, which enhances the perception and predictability of vehicular actions.

3. Agreement-seeking: A more advanced level, agreement-seeking, denotes the process of negotiating and establishing consensus among traffic participants for proposed actions, allowing for coordinated maneuvers.

4. Prescriptive: The most sophisticated level, prescriptive CDA, involves the vehicle adhering to specific directives such as traffic rules, control device states, or evacuation orders, where vehicle motion control becomes fully integrated with the broader traffic management system.



		SAE Driving Automation Levels					
		No Automation	**Driving Automation System**		**Automated Driving System (ADS)**		
		Level 0 *No Driving Automation (human does all driving)*	**Level 1** *Driver Assistance (longitudinal OR lateral vehicle motion control)*	**Level 2** *Partial Driving Automation (longitudinal AND lateral vehicle motion control)*	**Level 3** *Conditional Driving Automation*	**Level 4** *High Driving Automation*	**Level 5** *Full Driving Automation*
CDA Cooperation Classes	**No cooperative automation**	(e.g., Signage, TCD)	Relies on driver to complete the DDT and to supervise feature performance in real-time		Relies on ADS to perform complete DDT under defined conditions (fallback condition performance varies between levels)		
	Class A: Status-sharing *Here I am and what I see*	(e.g., Brake Lights, Traffic Signal)	Limited cooperation: Human is driving and must supervise CDA features (and may intervene at any time), and sensing capabilities may be limited compared to C-ADS		C-ADS has full authority to decide actions Improved C-ADS situational awareness beyond on-board sensing capabilities and increased awareness of C-ADS state by surrounding road users and road operators		
	Class B: Intent-sharing *This is what I plan to do*	(e.g., Turn Signal, Merge)	Limited cooperation (only longitudinal OR lateral intent that may be overridden by human)	Limited cooperation (both longitudinal AND lateral intent that may be overridden by human)	C-ADS has full authority to decide actions Improved C-ADS situational awareness through increased prediction reliability, and increased awareness of C-ADS plans by surrounding road users and road operators		
	Class C: Agreement-seeking *Let's do this together*	(e.g., Hand Signals, Merge)	N/A	N/A	C-ADS has full authority to decide actions Improved ability of C-ADS and transportation system to attain mutual goals by accepting or suggesting actions in coordination with surrounding road users and road operators		
	Class D: Prescriptive *I will do as directed*	(e.g., Hand Signals, Lane Assignment by Officials)	N/A	N/A	C-ADS has full authority to decide actions, except for very specific circumstances in which it is designed to accept and adhere to a prescriptive communication		

Figure 1.2: "Table 1 - Relationship between classes of CDA cooperation and levels of automation" in SAE J3216_202107 standard

The SAE J3216_202107 standard highlights that as we advance through the CDA cooperation classes, the "social" aspect of automated driving becomes increasingly dominant. This social nature implies that AVs or CAVs will not only be aware of each other's presence but also actively engage in communication, comprehension, negotiation, and collaboration with other road users and environmental elements. It is, therefore, necessary to study the social interactions between these entities to facilitate truly cooperative driving, thereby optimizing safety, efficiency, and user experience.

1.2 Thesis Overview

1.2.1 Research Objectives and Scope of Work

This dissertation focuses on developing a predictive control pipeline for CAVs in automated driving systems under mixed traffic conditions. Initially, we developed a deep learning-based lane change prediction model to accurately forecast the lane change behavior of surrounding vehicles. This model was designed to take into account the dynamic and unpredictable nature of lane change maneuvers, which are critical for preventing collisions and ensuring smooth traffic flow. Subsequently, we created a deep learning-based integrated two-dimensional trajectory prediction model that considers both lateral and longitudinal movements of vehicles. This model aims to provide more comprehensive and accurate predictions of vehicle trajectories by incorporating a wider range of motion patterns. Finally, we established a prediction-to-planning pipeline that utilizes the predicted outputs to provide safe and trustworthy planning results for CAVs in mixed traffic conditions. Specifically, the research objectives are as follows:

1. Develop vehicle motion prediction models that consider both historical trajectories and future interactions between vehicles to enhance prediction accuracy.

2. Create vehicle intention prediction models that account for the multimodality of driving behaviors, reflecting the diverse possible actions a driver might take in various traffic scenarios.

3. Integrate motion prediction with intention prediction in a probabilistic manner, allowing for the consideration of multiple potential future scenarios and their associated probabilities.

4. Implement multi-agent cooperative motion prediction and trajectory generation, enabling CAVs to anticipate and react to the movements of other vehicles in a coordinated manner.

5. Guide decision-making using interaction reasoning and motion planning-informed trajectory generation, ensuring that CAVs can make informed decisions based on a comprehensive understanding of their environment.

6. Optimize trajectories with a focus on safety and adherence to traffic rules, prioritizing the prevention of accidents and ensuring compliance with road regulations.

1.2.2 Research Contribution

This dissertation proposes an Integrated Framework of Motion Prediction and Planning for CAVs, comprising three main modules:

- `GIMTP`: **G**raph-based **I**nteraction-aware **M**ulti-modal **T**rajectory **P**rediction
- `GIRAFFE`: **G**raph-based **I**nteraction-aware **R**eliable **A**nticipative **F**easible **F**uture **E**stimator
- `RHINO`: **R**elational **H**ypergraph **I**nteraction-informed **N**eural m**O**tion generator and planner

- RHINO-Gen: Multi-agent Motion Generator
- RHINO-Plan: Neural Motion Planner

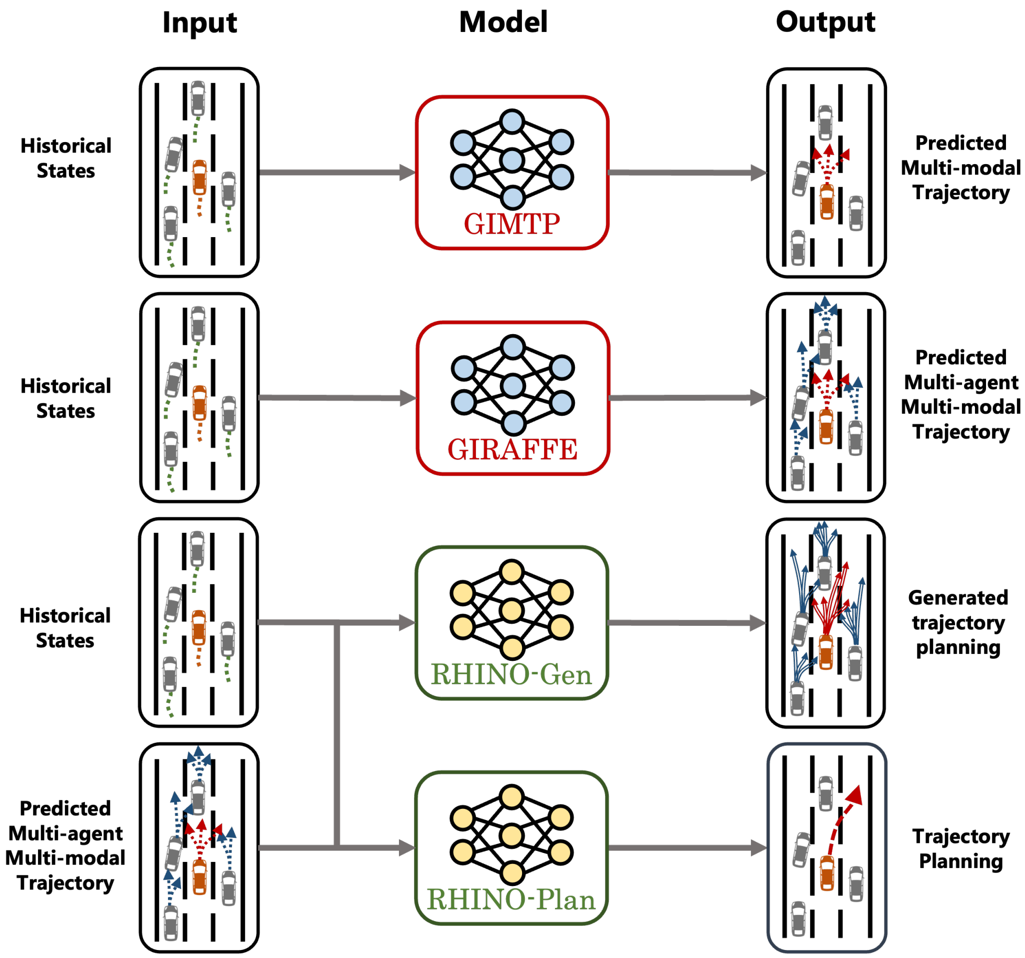

Input
Model
Output
Historical States
GIMTP
Predicted Multi-modal Trajectory
Historical States
GIRAFFE
Predicted Multi-agent Multi-modal Trajectory
Historical States
RHINO-Gen
Generated trajectory planning
Predicted Multi-agent Multi-modal Trajectory
RHINO-Plan
Trajectory Planning

Figure 1.3: Contributions.

As introduced in my prior work [29], the field of autonomous vehicle navigation has seen significant advancements with the introduction of the GIMTP framework. GIMTP marks a pivotal shift in how autonomous systems forecast vehicle trajectories, considering not only the vehicle's individual path but also the interwoven patterns of surrounding traffic. By utilizing a dynamic graph that encapsulates the nuanced motion states of vehicles, GIMTP sets the stage for understanding the sophisticated and bidirectional interactions that take place in real-time traffic scenarios. With an embedded system that leverages historical motion data, GIMTP provides a predictive glimpse into the future, illustrating potential vehicular interactions and their ripple effect on the road.

Building on the solid foundations of GIMTP, the upgraded GIRAFFE framework emerges as a robust Graph-based Interaction-aware Reliable Anticipative Feasible Future Estimator. GIRAFFE enhances GIMTP by integrating additional layers of predictive accuracy and anticipative capabilities. It leverages multi-modal and probabilistic techniques to not just predict but anticipate future states of traffic, accounting for a multitude of potential outcomes and their associated probabilities. This allows for a prediction mechanism that is both flexible and informed, capable of adjusting to the ever-changing tapestry of road dynamics. GIRAFFE's advancements represent a significant contribution to the autonomous vehicle field, offering a more reliable and comprehensive solution for multi-agent trajectory prediction, and paving the way for smarter, safer, and more socially adept autonomous driving systems.

RHINO, the Relational Hypergraph Interaction-informed Neural Motion generator and planner, represents the next frontier in motion generation and planning. It comprises two innovative components: RHINO-Gen and RHINO-Plan. RHINO-Gen, the Multi-agent Motion Generator, is an intricate system designed to synthesize motion states from a rich tapestry of group-wise interactions and behaviors observed in various agents. This system adeptly captures the social dynamics of driving, encoding the interactions within hypergraphs that mirror the complexities of real-world traffic situations. Through this representation, RHINO-Gen facilitates a deeper understanding of social behavior-inspired driving, enabling vehicles to engage in a more harmonious and socially aware navigation. RHINO-Plan, the Neural Motion Planner, builds on the foundations laid by RHINO-Gen. It utilizes the hypergraph-encoded information to craft strategic motion plans that consider not only the immediate future but also the extended horizon of vehicle interactions. By learning from the stochastic behaviors of agents, RHINO-Plan generates plausible trajectories that are iteratively refined, ensuring that the resulting motion plans prioritize safety, efficiency, and comfort. The RHINO framework, through its Gen and Plan components, exemplifies a significant leap toward actualizing human-like automated driving that is both intuitive and integrated into the societal norms of road sharing.

The main contributions of this research are summarized as follows:

1. Propose the GIMTP framework for forecasting vehicle motion.

2. Develop a dynamic graph to capture the evolving motion states of vehicles and to model the complex bidirectional and heterogeneous interactions occurring among them.

3. Map historical motion data onto future states through the application of an estimated future-guided graph embedding, highlighting the interdependen-

cies of future vehicular motion states.

4. Utilize a multi-modal and probabilistic approach for generating trajectory predictions by merging the feature space with potential semantic intentions and by forecasting potential future trajectories and their likelihoods.
5. Enhance `GIMTP` to `GIRAFFE` for improved multi-agent trajectory prediction capability.
6. Propose the `RHINO` framework for hypergraph-based interaction relational reasoning motion generation and planning.
7. Represent the group-wise interaction among different modalities of behaviors and the corresponding motion states of different agents through hypergraphs, enabling social behavior-inspired automated driving.
8. Incorporate interaction representation learning and relational reasoning to improve the social nature of automated driving.
9. Consider future relations and interactions and learn the posterior distribution to handle the stochasticity of each agent's behavior.
10. Generate plausible motion planning and refine the planned trajectory prioritizing safety, efficiency, and comfort, enabling human-like automated driving.

1.2.3 Organization of the Thesis

The thesis is systematically structured to present the research process. Chapter 1 introduces the topic by outlining the research subject, its importance, and the objectives. Chapter 2 offers a Literature Review, critically examining existing research to highlight knowledge gaps and establish the study's theoretical framework. Chapter 3 presents the Problem Statement, clearly defining the research challenges that will be addressed. The Methodology section, spread across multiple chapters, details the research design: Chapter 4 defines graphs and hypergraphs, Chapter 5 introduces the `GIMTP` model, and Chapter 6 covers the `RHINO` model. Chapter 7 thoroughly describes the Experiment Settings, specifying the framework and conditions under which the research was conducted. Chapter 8, Experiment Results and Discussions, interprets the findings in relation to the research questions and emphasizes the study's contributions to the field. The thesis concludes with Chapter 9, summarizing the findings and suggesting directions for future research.

Chapter 2

Literature Review

2.1 Interaction between Vehicles

Understanding interaction laws and their inherent complexities is crucial across a multitude of scientific and engineering fields, including autonomous driving [30], human behavior analysis [31], and interactions within chemical molecules [32] and brain networks [33]. Complex systems with interacting entities are typically modeled as graphs, where edges represent the corresponding interactions.

Understanding the interaction between vehicles is essential for predicting socially aware trajectories [17, 34]. Traditional methods, such as physics-based models [35] and the Kalman filter method [36], alongside classic machine learning approaches [37, 38, 39], often fall short in complex prediction contexts [16].

Recently, deep learning has become a favored tool for trajectory prediction due to its ability to learn intricate features and account for physics, road geometry, and vehicle interactions. Transforming historical information into representations that highlight temporal and spatial correlations is crucial. Techniques such as time series sequences [40, 41], occupancy grids [42, 43], and rasterized images [36, 44] have been utilized. Recurrent neural networks (RNN), particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) variants, effectively capture temporal correlations in sequential data. For spatial correlations, methods including Convolutional Neural Networks (CNN) [17, 42, 45, 46, 47, 48], attention-driven methods [49, 50, 51, 52, 53], and Graph Neural Networks (GNN) [54, 18, 55, 56, 57] have been proposed.

GNN-based methods are particularly promising for understanding non-Euclidean spatial dependencies, making them suitable for simulating interactions. However, most graph-based approaches rely on adjacency matrices based on neighborhood [58] and vehicular distances [59], which may not accurately reflect the influence among vehicles' motion states, especially on highways where longitudinal distances are more significant than lateral ones.

Furthermore, current methodologies often fail to address potential future interactions. Historical relationships do not inherently predict future interactions [60]. As illustrated in Figure 2.1, while the target vehicle and vehicle 2 show a weak relationship historically, they will interact significantly in the future. This highlights the challenge of predicting future states based solely on historical data. Accurate prediction requires considering future motion states of the target and surrounding vehicles, incorporating these influences into the trajectory prediction.

To enhance the representation of vehicle interactions and incorporate the influ-

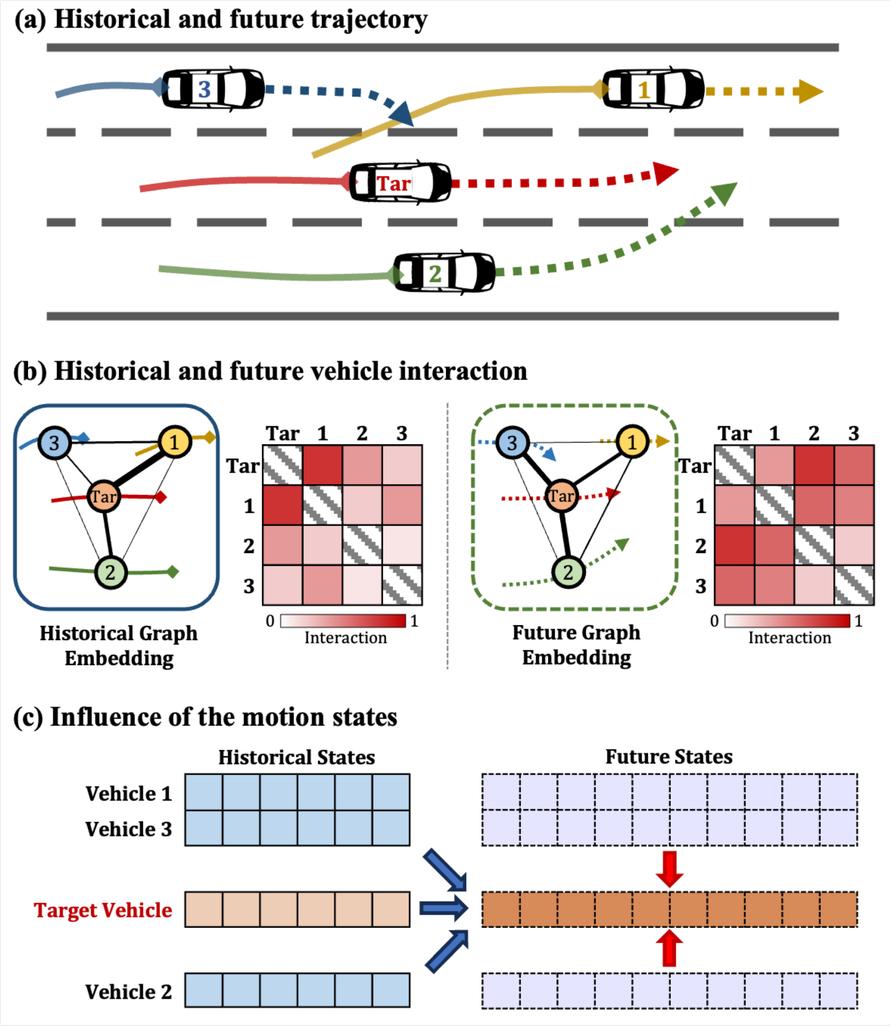

(a) Historical and future trajectory
3
1
Tar
2
(b) Historical and future vehicle interaction
Tar 1 2 3
Tar
1
2
3
0
1
Interaction
Historical Graph Embedding
Future Graph Embedding
(c) Influence of the motion states
Historical States
Future States
Vehicle 1
Vehicle 3
Target Vehicle
Vehicle 2

Figure 2.1: Insights of this research. Insight 1: Historical relationships do not inherently reflect future interactions. Insight 2: Surrounding vehicles' future motion states strongly influence the target vehicle.

ence of surrounding vehicles' motion states on the target vehicle, we propose constructing a graph with a dynamic adjacency matrix. This matrix captures interactions by considering elements such as neighborhood, distance, and potential risks derived from relative distances and velocities, assessing potential collision risks.

Utilizing forward and backward transition mechanisms in Diffusion Graph Convolutional Networks (DGCN), we capture bidirectional and heterogeneous interactions between the target vehicle and its surroundings. This technique integrates historical and predicted future motion states, creating comprehensive embeddings for refined future interaction estimation and seamless trajectory prediction integration.

2.2 Multi-modality of Driving Behaviors

The intricate interactions between vehicles result in unpredictable and uncertain traffic conditions, making it difficult to forecast a single accurate future trajectory [61]. Various factors, such as individual driver characteristics and psychological aspects, lead to different driving behaviors in identical situations [15, 62]. To account for the inherent multi-modality in driving behaviors, it is essential to generate multiple possible trajectories [13]. Models addressing this challenge can be categorized based on whether their latent variables have explicit semantics.

Driving intentions, which refer to the reasons behind a driver's movements and actions, are vital for ensuring safe and efficient traffic flow. In the first category, models use latent variables with clear semantics to represent these intentions. For example, studies by [15] and [42] utilize driving maneuvers as latent variables to capture various behaviors. [40] employs attention mechanisms for trajectory generation within specific scene contexts. These maneuvers are categorized and integrated into multi-modal trajectory prediction models using Gaussian distributions. Some models opt for pre-clustered anchors instead of predefined maneuvers, predicting goals with vector anchors and high-definition maps [63, 64, 65, 66], while others use diverse lane features for predictions [67]. However, many studies rely on simple concatenation operations to merge maneuver or lane features with encoded context, which is insufficient for producing diverse predictions.

The second category involves models where latent variables lack explicit semantics, benefiting from advances in generative deep learning, particularly Variational Autoencoders (VAE) [68, 69] and Generative Adversarial Networks (GAN) [70, 47, 71]. GANs integrate diverse factors in their generators [72, 73, 74], with discriminators assessing latent details. VAEs use encoder-decoder structures to enhance multi-modal trajectory predictions [75, 76]. These models employ latent random variables to create varied, multi-modal trajectories by adding noise from latent distributions to encoded features, resulting in stochastic outcomes. Challenges with these approaches include a lack of interpretability and difficulties in determining optimal sample sizes and probabilistic assignments for trajectories.

This work aims to improve interpretability by treating multi-modal behaviors as latent variables with explicit semantics, effectively representing potential driving intentions. Building on the approach of [50], we propose a probabilistic framework to predict semantic intentions at each future timestamp. This framework connects intrinsic intentions with feasible trajectories through intention-specific feature combinations, achieving comprehensive multi-modal prediction.

2.3 Relational Learning and Reasoning

Relational reasoning has emerged as a crucial aspect of AI, especially within multi-agent systems operating in shared environments. This capability to comprehend, interpret, and forecast interactions between entities is vital for the development of intelligent systems. This review explores multi-agent relational reasoning and its application in vehicle trajectory prediction and planning, emphasizing current challenges and prospective advancements.

In multi-agent systems, relational reasoning involves deciphering the intricate interactions among agents and their surroundings, which is essential for coordinated actions, accurate outcome prediction, and informed decision-making. For autonomous vehicles and intelligent transportation systems, relational reasoning is critical in predicting the trajectories of interacting agents such as vehicles, pedestrians, and cyclists, forming the basis for safe and efficient trajectory planning algorithms.

Traditional model-based approaches laid the groundwork by creating mathematical models of vehicle dynamics and interactions. However, these models often lacked the flexibility and scalability needed for dynamic environments. Early techniques like Locally Linear Embedding (LLE) [77] and Isomap [78] were effective with limited data but faced performance constraints. The advent of deep learning has transformed vehicle trajectory prediction by utilizing large datasets to capture complex spatial and temporal dependencies, although these methods can be opaque and difficult to interpret.

Deep learning approaches model interactions through social operations [17, 71], graph-based modeling [18, 55, 56], and attention mechanisms [50, 51, 52]. While these methods have enhanced trajectory prediction, they often fall short in relational reasoning, primarily focusing on pairwise relationships and struggling with the limited observable information in multi-agent systems. Some works, such as NRI [79] and EvolveGraph [21], have made strides in explicit relational reasoning but are restricted to pairwise interactions and interaction categories without considering interaction intensity.

It is crucial to study group-wise behaviors beyond pairwise interactions. In multi-agent systems, dynamic interactions among agents frequently exhibit cooperative and competitive behaviors [11, 12, 24]. These group-level interactions are common but have been rarely modeled. Traditional graph-based methods, though effective at capturing pairwise interactions, fall short in representing the collective influence of groups of entities on each other's behaviors in real-world traffic scenarios. There is also a need to model the group-wise interactions of multi-modal driving behaviors among multiple agents. In dense traffic scenarios, vehicles often change lanes, accelerate, decelerate, and engage in complex interactions with other agents. These interactions involve multiple possible longitudinal and lateral behaviors of multiple entities influencing each other simultaneously, which cannot be fully captured by considering only pairwise relationships. Additionally, the unpredictable nature of human behavior adds further complexity, necessitating models capable of handling the uncertainty and variability in agent behaviors, which is crucial for safe and efficient driving. Therefore, a significant challenge in achieving socially inspired automated driving is representing and reasoning about the diverse interactions among agents and their multiple possible behaviors [20, 25, 26].

Advancements include “EvolveGraph” [21], which dynamically constructs and updates a graph representing agent relationships, capturing temporal interaction changes for a nuanced understanding of multi-agent environments. “EvolveHypergraph” uses hypergraphs to represent higher-order relationships, providing a richer representation of relational dynamics. “GroupNet” [20] and “DynGroupNet” [80] model group behaviors, with DynGroupNet incorporating dynamic group formation and dissolution to reflect real-world fluid agent interactions.

Despite these advancements, challenges remain, including scalability with interaction complexity, agent heterogeneity, and non-stationary behavior patterns, which complicate modeling and prediction. Ensuring safety and reliability under uncertainty and incomplete information, along with integrating ethical and legal considerations, are ongoing issues in motion prediction and planning.

To advance relational reasoning in trajectory prediction, this work focuses on capturing and representing interactions. Building on the framework of [20], we propose a multiscale hypergraph for modeling group-wise interactions of varying sizes, learned in a data-driven manner rather than being handcrafted. For interaction embedding, we introduce a three-element representation format: neural interaction strength, category, and per-category function, capturing the interaction strength and category in interactive groups. Neural message passing over the multiscale hypergraph integrates this interaction embedding into the representation learning process.

2.4 Learning-based Motion Planning

Motion planning is closely tied to motion prediction and has evolved significantly through methods such as path optimization [81] and sampling [82]. However, safely navigating complex and interactive traffic environments requires incorporating the predicted behaviors of other participants [83]. Learning-based motion planning [84, 85] has garnered attention for its ability to manage a wide range of driving scenarios, achieving notable results with unified neural networks [86]. For instance, PiP [87] iteratively makes conditional predictions to adjust sampling-based planning, though it only considers marginal futures and is limited by the planning paths it generates. DIPP [88] merges differentiable planning objectives with joint trajectory predictions, allowing for more responsive planning. Despite the impressive performance of learning-based motion planning in diverse scenarios, its robustness and safety can be affected by instability.

Many existing learning-based motion planning methods incorporate predictions by using the future trajectories of surrounding actors. However, they often overlook the group-wise interactions among multiple potential future motion states of several agents. The three main challenges (Figure 1.1) in motion planning remain paramount. First, precise motion planning depends on accurately predicting the multi-modality of driving behaviors, which requires robust interaction models capable of forecasting multiple potential trajectories [89]. Second, the dynamic interactions between multiple agents introduce significant uncertainty, necessitating models that can adaptively respond to varying traffic conditions and behaviors. Third, relational reasoning about interactions is critical in motion planning, as it involves not only predicting potential trajectories but also understanding the diverse interactions among agents to make informed and safe planning decisions.

To tackle these challenges, we propose integrating prediction guidance for motion planning by learning group-wise interaction patterns and potential future motion states based on multi-agent multi-modal predictions. This approach aims to develop a planning system that is safer, more robust, and socially compliant. By focusing on group-wise interactions and adapting to the dynamic and uncertain nature of traffic environments, this method strives to enhance the overall safety and efficiency of motion planning in autonomous driving systems.

Chapter 3

Problem Statement

3.1 Problem Statement

This study presents an end-to-end framework designed to provide motion trajectory predictions for each target vehicle and its surrounding vehicles, as well as motion planning for the target vehicle in a multi-lane highway scenario. The target vehicle is assumed to be a connected and autonomous vehicle (CAV), while the surrounding vehicles can include both human-driven vehicles and other CAVs. The prediction component aims to probabilistically forecast the multi-modal trajectories of the target and surrounding vehicles by leveraging their historical motion states and interactions. Meanwhile, the planning component focuses on developing a planned trajectory for the target vehicle by utilizing the predicted multimodal motion states of both the target and its surrounding vehicles, while also reasoning about their interactions.

Our research focuses on several key aspects: analyzing the interactions between the target vehicle and its surrounding vehicles, understanding the multi-modality of driving behaviors, and exploring the relationship between multi-modal future motion states and historical motion states. This comprehensive approach aims to enhance the accuracy and safety of trajectory prediction and motion planning in complex traffic environments.

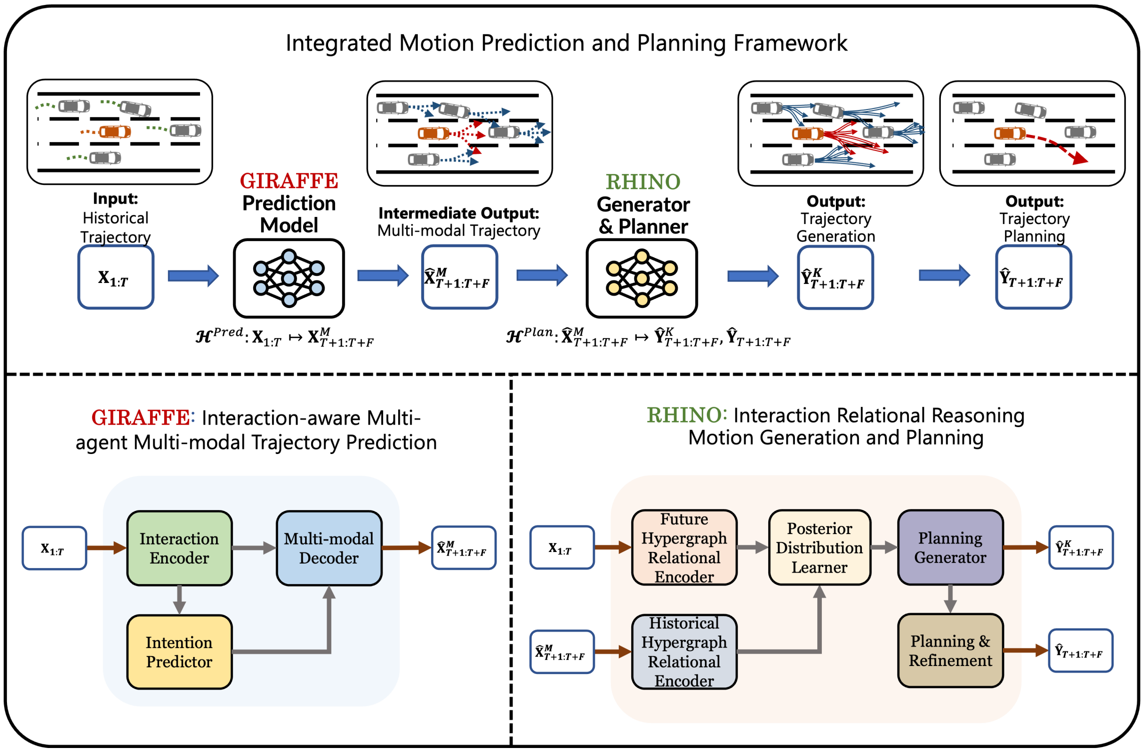

Integrated Motion Prediction and Planning Framework
Input: Historical Trajectory
$\mathbf{X}_{1:T}$
GIRAFFE Prediction Model
$\mathcal{H}^{Pred}: \mathbf{X}_{1:T} \mapsto \mathbf{X}^{M}_{T+1:T+F}$
Intermediate Output: Multi-modal Trajectory
$\hat{\mathbf{X}}^{M}_{T+1:T+F}$
RHINO Generator & Planner
$\mathcal{H}^{Plan}: \hat{\mathbf{X}}^{M}_{T+1:T+F} \mapsto \hat{\mathbf{Y}}^{K}_{T+1:T+F}, \hat{\mathbf{Y}}_{T+1:T+F}$
Output: Trajectory Generation
$\hat{\mathbf{Y}}^{K}_{T+1:T+F}$
Output: Trajectory Planning
$\hat{\mathbf{Y}}_{T+1:T+F}$
GIRAFFE: Interaction-aware Multi-agent Multi-modal Trajectory Prediction
$\mathbf{X}_{1:T}$
Interaction Encoder
Multi-modal Decoder
Intention Predictor
$\hat{\mathbf{X}}^{M}_{T+1:T+F}$
RHINO: Interaction Relational Reasoning Motion Generation and Planning
$\mathbf{X}_{1:T}$
$\hat{\mathbf{X}}^{M}_{T+1:T+F}$
Future Hypergraph Relational Encoder
Historical Hypergraph Relational Encoder
Posterior Distribution Learner
Planning Generator
Planning & Refinement
$\hat{\mathbf{Y}}^{K}_{T+1:T+F}$
$\hat{\mathbf{Y}}_{T+1:T+F}$

Figure 3.1: Framework architecture.

3.2 Model Input

Mathematically, the task of vehicle trajectory prediction can be formulated as predicting the probability distribution of the target vehicle's future trajectory position based on the observed historical motion information of the target vehicle and its surrounding vehicles. The historical states of the vehicle group over a historical time horizon $[1, \ldots, T]$ can be denoted as $X_{1:T} = \{X_1, X_2, \ldots, X_T\}$. Each historical state X_t at time step $t \in \{1, \ldots, T\}$ represents the union set of the historical states of the target vehicle and its surrounding vehicles. Thus, $X_t = \{\mathbf{x}_t^0, \mathbf{x}_t^1, \ldots, \mathbf{x}_t^N\}$, where $\mathbf{x}_t^i$ represents the historical state of vehicle i, for all $i \in \{0, 1, \ldots, N\}$ and all $t \in \{1, \ldots, T\}$. In this thesis, superscripts refer to vehicle indices, with $i = 0$ specifically for the target vehicle, and the subscript referring to time steps. The state $\mathbf{x}_t^i$ associated with vehicle i could include its longitudinal and lateral positions and velocity.

3.3 Model Output

3.3.1 Single Agent Motion Prediction

Assume the current time step is T. The predicted states of the vehicle group for a future time horizon $[T+1, \ldots, T+F]$ are denoted as $\hat{Y}_{T+1:T+F} = \{\hat{Y}_{T+1}, \hat{Y}_{T+2}, \ldots, \hat{Y}_{T+F}\}$. Each predicted future state $\hat{Y}_{T+f}$ at time step $T+f$ only includes the predicted state of the target vehicle. Thus, $\hat{Y}_{T+f} = \{\hat{\mathbf{y}}_{T+f}^0\}$, where $\hat{\mathbf{y}}_{T+f}^0$ denotes the future state of the target vehicle at time step $T + f$ for all $f \in \{1, \ldots, F\}$. The state $\hat{\mathbf{y}}_{T+f}^0$ encompasses the longitudinal and lateral positions of the target vehicle at time step $T+f$.

It is important to note that the coordinates of all vehicles in the vehicle group are expressed in a reference frame where the origin is the position of the target vehicle at timestamp T. The model input consists of $X_{1:T} = \{X_1, X_2, \ldots, X_T\}$ during the past T time steps, and the output of the model is a probability distribution $P(\hat{Y}_{T+1:T+F}|X_{1:T})$ over the next F time steps. In this work, the distribution of $\mathbf{y}^0_{T+f}$ is parameterized as a bivariate Gaussian distribution with mean $(\mu_{T+f,x}, \mu_{T+f,y})$, variance $(\sigma^2_{T+f,x}, \sigma^2_{T+f,y})$, and correlation coefficient ρ_{T+f}, where the subscript x stands for the longitudinal position and y stands for the lateral position.

3.3.2 Multi-Agent Motion Prediction

For multi-agent multi-modal motion prediction, the input to the model consists of $X_{1:T} = \{X_1, X_2, \ldots, X_T\}$ during the past T time steps, and the output of the model is the estimated future trajectories $\hat{Y}^M_{T+1:T+F}$ for the next F time steps. The predicted states of the vehicle group in a future time horizon $[T+1, \ldots, T+F]$ can be denoted as $\hat{Y}^M_{T+1:T+F} = \{\hat{Y}^M_{T+1}, \hat{Y}^M_{T+2}, \ldots, \hat{Y}^M_{T+F}\}$. Each predicted future state $\hat{Y}^M_{T+f}$ at time step $T+f$ is composed of the predicted motion states with M modalities of all the vehicles in the vehicle group. Thus $\hat{Y}^M_{T+f} = \{\hat{\mathbf{y}}^{i,m}_{T+f}\}$, where $\hat{\mathbf{y}}^{i,m}_{T+f}$ denotes the future states of each vehicle i in the vehicle group with M modalities at time step $T+f$ for all $i \in \{0, 1, \ldots, N\}$, all $m \in \{1, \ldots, M\}$, and all $f \in \{1, \ldots, F\}$. The state $\hat{\mathbf{y}}^0_{T+f}$ consists of the longitudinal and lateral positions of the target vehicle at time step $T+f$.

3.3.3 Motion Planning

In the context of motion planning, the input to the model includes the historical states of the vehicle group $X_{1:T} = \{X_1, X_2, \ldots, X_T\}$ during the past T time steps and $\hat{Y}^M_{T+1:T+F} = \{\hat{Y}^M_{T+1}, \hat{Y}^M_{T+2}, \ldots, \hat{Y}^M_{T+F}\}$, which is the estimated multi-modal future trajectories of all the vehicles in the vehicle group. The output of the model is the planned trajectories $\hat{Y}_{T+1:T+F}$ over the next F time steps. The predicted states of the vehicle group in a future time horizon $[T+1, \ldots, T+F]$ can be denoted as $\hat{Y}_{T+1:T+F} = \{\hat{Y}_{T+1}, \hat{Y}_{T+2}, \ldots, \hat{Y}_{T+F}\}$. Each predicted planned state $\hat{Y}_{T+f}$ at time step $T+f$ is composed of the planned trajectory of the target vehicle. Thus $\hat{Y}_{T+f} = \{\hat{\mathbf{y}}^0_{T+f}\}$, where $\hat{\mathbf{y}}^i_{T+f}$ denotes the planned trajectory of the target vehicle at time step $T+f$ for all $f \in \{1, \ldots, F\}$. The state $\hat{\mathbf{y}}^0_{T+f}$ consists of the longitudinal and lateral positions of the target vehicle at time step $T+f$.

3.4 Problem Statement in Mathematical Formulation

The trajectory prediction problem can be summarized as follows: Given the states $X_{1:T}$ of all the vehicles in a vehicle group over a past time horizon $[1, \ldots, T]$, the objective is to train a model $\mathbf{H}^{\mathbf{Pred}}(\cdot)$ to predict the trajectory distributions $\hat{Y}_{T+1:T+F}$ of the target vehicle that approximate the ground truth trajectory Y in the future time horizon $[T+1, \ldots, T+F]$:

$$\hat{X}^M_{T+1:T+F} = \mathbf{H}^{\mathbf{Pred}}(X_{1:T}) \tag{3.1}$$

The subsequent trajectory planning problem can be summarized as follows: Given the states $X_{1:T}$ of all the vehicles in a vehicle group over a past time horizon

$[1, \ldots, T]$ and the estimated multi-modal future motion states of the vehicles in the vehicle group, the goal is to train a neural planner model $\mathbf{H}^{\mathbf{Plan}}(\cdot)$ to generate a planned future trajectory $\hat{Y}_{T+1+F}$ of the target vehicle in the future time horizon $[T+1, \ldots, T+F]$ considering safety, efficiency, and comfort:

$$\hat{Y}_{T+1:T+F} = \mathbf{H}^{\mathbf{Plan}}(X_{1:T}, \hat{X}^{M}_{T+1:T+F}) \tag{3.2}$$

3.5 Scenario Settings

Our scenario unfolds on a multi-lane highway characterized by a mixed traffic environment where CAVs coexist with human-driven vehicles (HDVs). The artery of our smart transportation network is lined with sophisticated Roadside Units (RSUs), serving as beacons of communication and data gathering. These elements converge to create an ecosystem where information flows seamlessly between vehicles and infrastructure, orchestrating a symphony of coordinated movement aimed at enhancing safety and optimizing traffic flow. The intelligence levels of automated vehicles follow the guidelines in [90, 91, 28], and the intelligence levels of roadside infrastructure are designed following [92, 91].

In the first deployment phase, we introduce CAVs equipped with intelligence levels ranging from L3 to L4.. These vehicles represent a significant step towards full automation, capable of self-driving in specific scenarios with little to no human intervention. These CAVs are outfitted with advanced sensors that meticulously collect data from their local traffic environment, enabling them to navigate and respond to on-the-road situations with precision. Furthermore, these CAVs are not isolated units; they share the acquired data with fellow CAVs, creating a network of shared awareness that ensures individual and collective safety and efficiency.

The second deployment phase sees the integration of CAVs operating at a slightly lower spectrum of autonomy, within the L1 to L3 intelligence bracket, coupled with the support of RSUs ranging from intelligence levels I2 to I4. In this arrangement, the CAVs continue to gather local traffic data and communicate with each other. The RSUs augment this system by collecting comprehensive data on regional traffic conditions, thus providing a macroscopic view of the traffic landscape. The RSUs serve a dual role, not only as data collectors but also as vital nodes that relay this information back to the CAVs. This partnership enhances the capabilities of CAVs, especially those that still rely on a degree of human control, ensuring a safer and more informed vehicular operation within the highway environment.

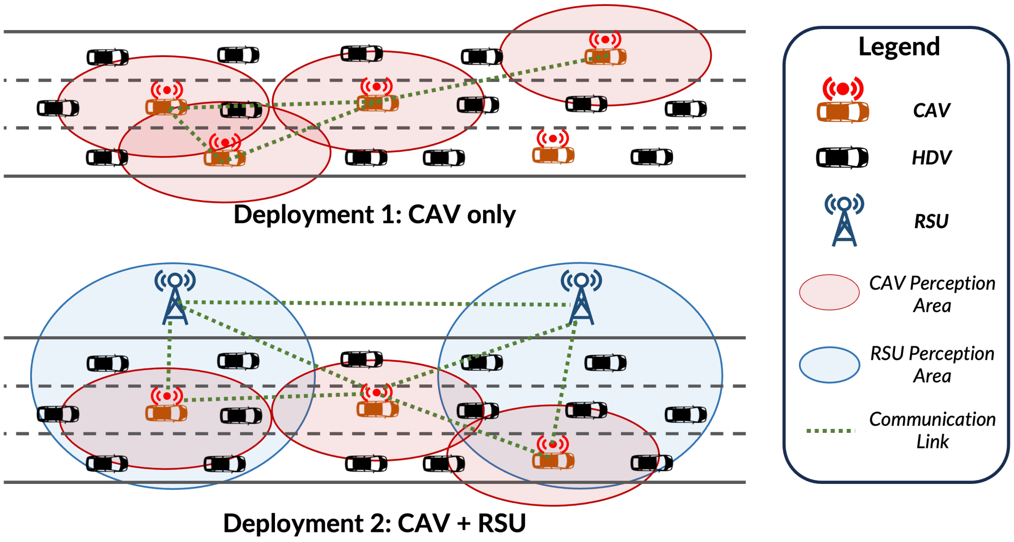


Figure 3.2: Study scenarios.

3.6 Assumptions

3.6.1 Assumptions for Communication and Sensing

In the realm of communication, the network relies on Vehicle-to-Everything (V2X) technology [93, 94, 95, 96, 97], which is critical for the functioning of connected vehicles. The system's range is impressive, extending approximately 300 to 600 meters, enabling vehicles to communicate over considerable distances. For the purpose of our study and to simplify the modeling, we will assume that there is no communication delay.

The sensor suite on each vehicle consists of radar [98, 99], camera [100, 101, 102], and LiDAR [103, 104] systems, each with its own specific detection range. Radar can detect objects at a range of about 150 to 250 meters, whereas cameras can observe the environment at a distance of 150 to 300 meters. LiDAR, which stands for Light Detection and Ranging, boasts the most extensive range of approximately 200 to 600 meters. These ranges are the backbone for the longitudinal study range, defined by the average minimax sensor detection range, and a spacing parameter set to span from -100 to +150 meters.

3.6.2 Study Range Parameters

When considering the study range for this project, two main dimensions are taken into account: longitudinal and lateral ranges. As shown in Fig. 3.3, longitudinally, the study accounts for an average of the minimum and maximum sensor detection ranges, incorporating a spacing that allows for measurements ranging approximately from -300 to +500 feet relative to the target vehicle. Laterally, the focus is narrowed to one adjacent lane on each side, measured from the centroid of the target vehicle to the outer lane markings of adjacent lanes, a distance of roughly ±5.5 meters or ±18 feet.

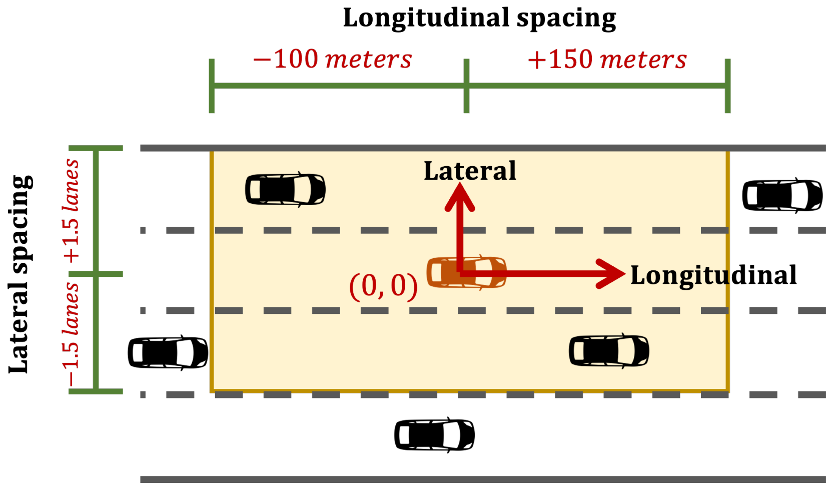


Figure 3.3: Study range assumptions.

3.6.3 Study Object Focus

The study treats each individual vehicle as a potential target vehicle, as shown in Figure 3.4. This includes the acquisition of historical motion states data of the target vehicle itself, as well as the surrounding vehicles, both CAVs and HDVs. A key assumption is that the driving behaviors are independent and identically distributed (i.i.d.), ensuring that the data collected provides a consistent basis for analysis.

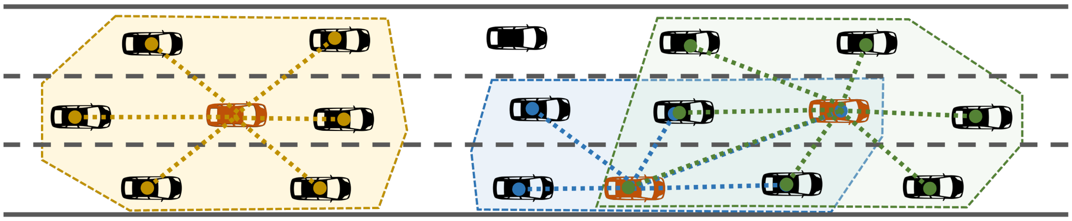


Figure 3.4: Study object assumptions.

3.7 Objective of the Study

The ultimate goal is to provide accurate motion prediction for all vehicles on the highway. By understanding historical motion states and sensing ranges, the system aims to aid in decision-making, motion planning, and control processes for CAVs. The sophisticated mesh of communication and detection technologies is crucial for this objective, as it allows for a comprehensive understanding of vehicular movement and behavior, which is essential for the development of reliable automated systems on our roads.

3.8 Framework Architecture

The proposed framework adopts an end-to-end architecture, as shown in Figure 3.5, which involves three major components:

1. `GIMTP`: **G**raph-based **I**nteraction-aware **M**ulti-modal **T**rajectory **P**rediction, which provides multi-modal trajectory prediction for the target agent by modeling the historical and future interactions as graphs.

2. GIRAFFE: **G**raph-based **I**nteraction-aware **R**eliable **A**nticipative **F**easible **F**uture **E**stimator, as an enhanced version of GIMTP, which provides multi-modal trajectory prediction for multiple agents by modeling the historical and future interactions as graphs.

3. RHINO: **R**elational **H**ypergraph **I**nteraction-informed **N**eural m**O**tion generator and planner, which provides motion generation and motion planning by modeling the interaction relation using multiscale hypergraph representations.

Here, the overall framework architecture adopts GIRAFFE and RHINO as two major components to fulfill the integrated motion prediction and planning functions.

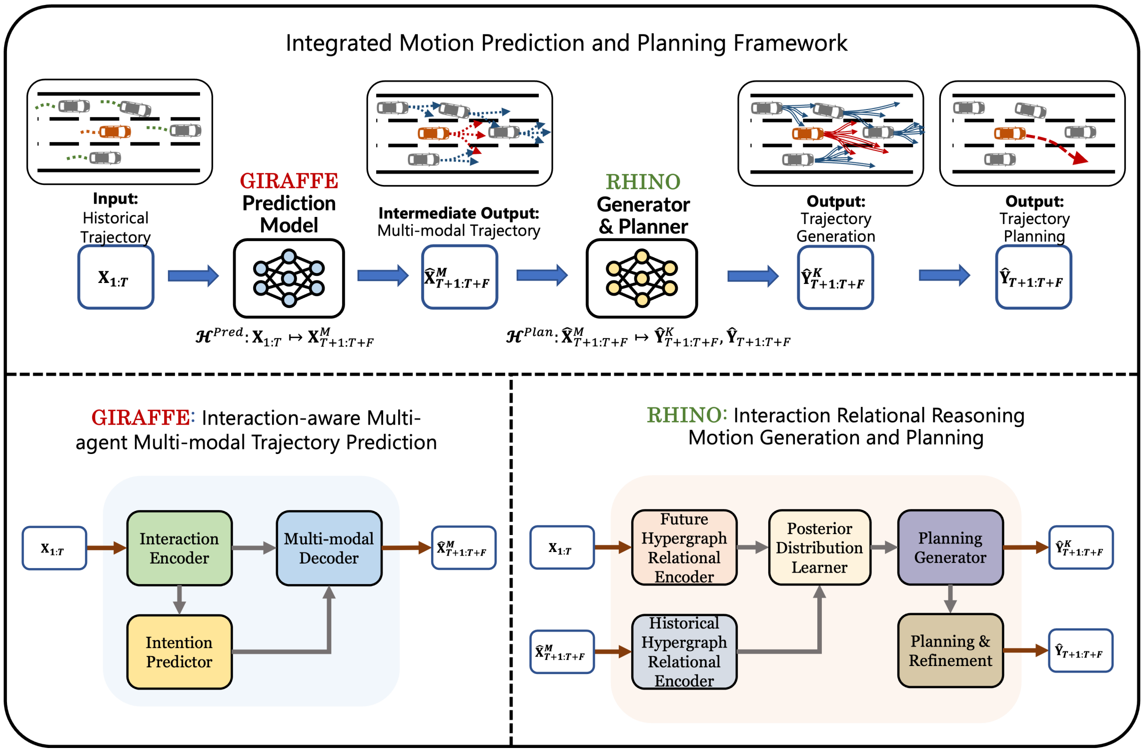

Integrated Motion Prediction and Planning Framework
Input:
Historical Trajectory
$\mathbf{X}_{1:T}$
GIRAFFE
Prediction Model
$\mathcal{H}^{Pred}: \mathbf{X}_{1:T} \mapsto \mathbf{X}^{M}_{T+1:T+F}$
Intermediate Output:
Multi-modal Trajectory
$\hat{\mathbf{X}}^{M}_{T+1:T+F}$
RHINO
Generator & Planner
$\mathcal{H}^{Plan}: \hat{\mathbf{X}}^{M}_{T+1:T+F} \mapsto \hat{\mathbf{Y}}^{K}_{T+1:T+F}, \hat{\mathbf{Y}}_{T+1:T+F}$
Output:
Trajectory Generation
$\hat{\mathbf{Y}}^{K}_{T+1:T+F}$
Output:
Trajectory Planning
$\hat{\mathbf{Y}}_{T+1:T+F}$
GIRAFFE: Interaction-aware Multi-agent Multi-modal Trajectory Prediction
$\mathbf{X}_{1:T}$
Interaction Encoder
Multi-modal Decoder
Intention Predictor
$\hat{\mathbf{X}}^{M}_{T+1:T+F}$
RHINO: Interaction Relational Reasoning Motion Generation and Planning
$\mathbf{X}_{1:T}$
$\hat{\mathbf{X}}^{M}_{T+1:T+F}$
Future Hypergraph Relational Encoder
Historical Hypergraph Relational Encoder
Posterior Distribution Learner
Planning Generator
Planning & Refinement
$\hat{\mathbf{Y}}^{K}_{T+1:T+F}$
$\hat{\mathbf{Y}}_{T+1:T+F}$

Figure 3.5: Framework architecture.

Chapter 4

Definitions of Graphs and Hypergraphs

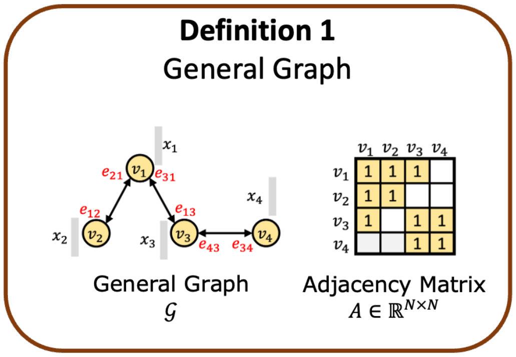
4.1 Graph Definitions

Graphs are excellent representations to describe and analyze entities with relations and interactions. Specifically, in a multi-agent system, a graph representation is used by modeling each agent as a node and the pair-wise interaction as the edge. Thus, we define the graph as follows:

Definition 1 (Graph) *A graph $\mathcal{G}$ is a representation describing and analyzing entities with relations and interactions, which can be represented by a set of nodes and a set of edges that establish relationships between these nodes. The graph $\mathcal{G}$ is expressed as*

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}; X, A)$$

where $\mathcal{V} \in \mathbb{R}^N$ denotes the node set, $\mathcal{E} \in \mathbb{R}^{|N \times N|}$ denotes the edge set, $X \in \mathbb{R}^{N \times C}$ represents the feature tensor, and $A \in \mathbb{R}^{N \times N}$ indicates the adjacency matrix.



	v_1	v_2	v_3	v_4
v_1	1	1	1	
v_2	1	1		
v_3	1		1	1
v_4			1	1

Figure 4.1: General Graph.

By modeling each vehicle agent as a node in a graph, we can further define an Agent Graph with the same structure as a general graph:

Definition 2 (**Agent Graph**) *Let $\mathcal{G}^a$ be a graph representation of the motion states and interaction of N agents, with each agent represented as a node. $\mathcal{G}^a$ is expressed as*

$$\mathcal{G}^a = (\mathcal{V}^a, \mathcal{E}^a; X^a, A^a)$$

where $\mathcal{V}^a \in \mathbb{R}^N$ denotes the node set, $\mathcal{E}^a \in \mathbb{R}^{|N \times N|}$ denotes the edge set, $X^a \in \mathbb{R}^{N \times C}$ represents the feature tensor, and $A^a \in \mathbb{R}^{N \times N}$ indicates the adjacency matrix.

In our previous work [29] and the enhanced version `GIRAFFE`, we model and represent the motion states and the interactions between the agents as Agent Graphs.

To better represent the interactions and relations of the predicted multi-agent multi-modal motion states with graphs, we can expand each agent node with the

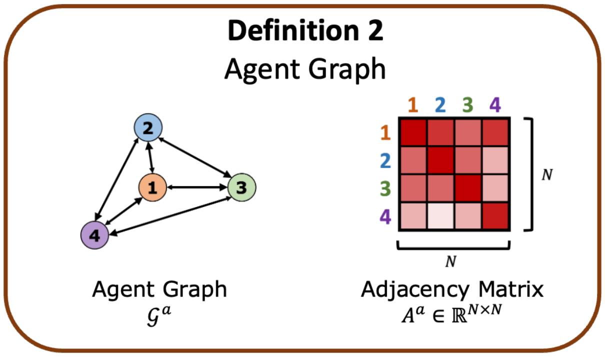


Figure 4.2: Agent Graph.

number of modalities of the behavior. Thus, we can construct an Agent-Behavior Graph.

Definition 3 (Agent-Behavior Graph) *Let $\mathcal{G}^b$ be a graph representation of the multi-modal motion states of N agents, with each of M behavior modes for each agent represented as a node. $\mathcal{G}^b$ is expressed as*

$$\mathcal{G}^b = (\mathcal{V}^b, \mathcal{E}^b; X^b, A^b)$$

where $\mathcal{V}^b \in \mathbb{R}^{|MN|}$ denotes the node set, $\mathcal{E}^b \in \mathbb{R}^{|MN\times MN|}$ denotes the edge set, $X^b \in \mathbb{R}^{|MN|\times C}$ represents the feature tensor, and $A^b \in \mathbb{R}^{|MN|\times|MN|}$ indicates the adjacency matrix.

4.2 Introduction to Hypergraphs

In the realm of autonomous and connected vehicular traffic systems, there are fundamental facts and challenges that underscore the necessity for advanced interaction representation. Among these, the inherent complexity of vehicular in-

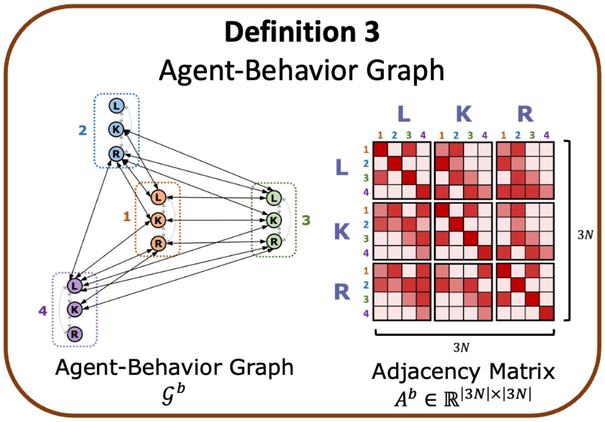


Figure 4.3: Agent-Behavior Graph.

teractions can be distilled into two primary categories: pair-wise interaction and group-wise interaction, each presenting unique difficulties and opportunities for improving overall traffic dynamics.

Pair-wise interaction represents the fundamental unit of vehicular interaction, focusing on the dyadic relationships between two vehicles. This dimension is characterized by direct vehicle-to-vehicle communication and has significant implications on the scale of interaction, where decisions made by one vehicle can have an immediate and localized impact on another. The challenges here involve ensuring that decision-making between pairs of vehicles is optimized to promote traffic flow efficiency and safety. The scalability and flexibility of pair-wise interaction models are also crucial to adapt to various traffic conditions and to incorporate an expanding number of vehicle types and communication technologies.

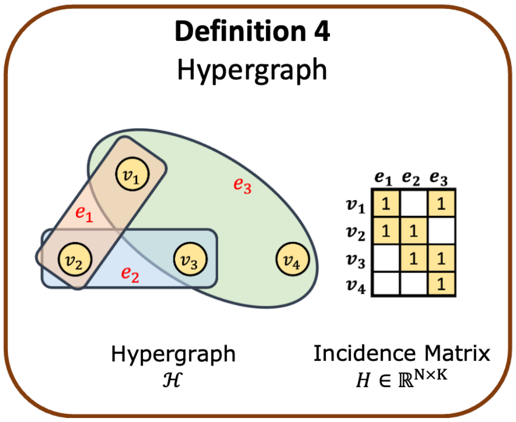
On the other hand, group-wise interaction expands this focus to consider the collective behavior of clusters of vehicles. Here, the challenges magnify as the complexity of interactions increases with the number of vehicles involved. It is not merely the sum of pair-wise interactions but also includes the emergent behaviors that arise from group dynamics. These can manifest in the form of platooning, cooperative lane changes, and synchronized maneuvers that aim to enhance traffic flow efficiency at a larger scale. Safety within group-wise interactions becomes even more critical, as a single misstep could have cascading effects throughout the vehicle cluster.

The conclusion drawn from analyzing these interactions points towards the necessity of a hypergraph interaction representation. Unlike traditional graph models that limit relationships to pairs, hypergraphs allow for a more nuanced and comprehensive representation of group-wise dynamics. By modeling the connections among multiple vehicles simultaneously, hypergraphs provide a more robust framework for understanding and optimizing the intricate web of interactions that define modern traffic systems. This holistic approach is anticipated to yield substantial advancements in traffic management, vehicle routing algorithms, and overall transportation system design.

Definition 4 (Hypergraph) *A hypergraph $\mathcal{H}$ is a natural extension of general graphs, allowing an edge to join any number of nodes, which can represent the higher-order relationships involving multiple entities. The hypergraph $\mathcal{H}$ is expressed as*

$$\mathcal{H} = (\mathcal{V}, \mathcal{E}; X, H)$$

where $\mathcal{V} \in \mathbb{R}^N$ denotes the node set, $\mathcal{E} \in \mathbb{R}^K$ denotes the edge set (K can be complex), $X \in \mathbb{R}^{N \times C}$ represents the feature tensor, and $H \in \mathbb{R}^{N \times K}$ indicates the incidence matrix, *where H_{ij} indicates whether node v_i is part of the hyperedge e_j.*



	e_1	e_2	e_3
v_1	1		1
v_2	1	1	
v_3		1	1
v_4			1

Figure 4.4: Hypergraph.

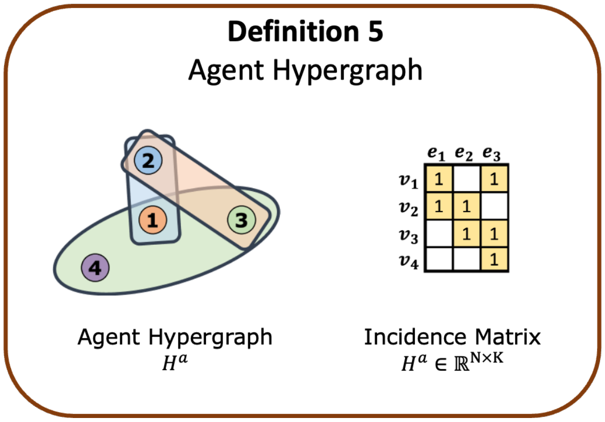
Similarly, we can define an Agent Hypergraph with the same structure as a hypergraph to represent the agents and group-wise interactions among them:

Definition 5 (Agent Hypergraph) *Let $\mathcal{H}^a$ be a hypergraph representation of the motion states of N agents, with each agent represented as a node. The hypergraph $\mathcal{H}^a$ is expressed as*

$$\mathcal{H}^a = (\mathcal{V}^a, \mathcal{E}^a; X^a, H^a)$$

where $\mathcal{V}^a \in \mathbb{R}^N$ denotes the node set, $\mathcal{E}^a \in \mathbb{R}^K$ denotes the edge set (K can be complex), $X^a \in \mathbb{R}^{N \times C}$ represents the feature tensor, and $H^a \in \mathbb{R}^{N \times K}$ indicates the incidence matrix, where H^a_{ij} indicates whether node v_i is part of the hyperedge e_j.

The intricate nature of vehicular dynamics on multi-lane highways necessitates an understanding that extends beyond individual behaviors to encompass the collective. This nature delves into the realm of group-wise interaction, which



$$\begin{array}{c|ccc} & e_1 & e_2 & e_3 \\ \hline v_1 & 1 & & 1 \\ v_2 & 1 & 1 & \\ v_3 & & 1 & 1 \\ v_4 & & & 1 \end{array}$$

Figure 4.5: Agent Hypergraph.

involves the multifaceted behaviors of multiple vehicles. Within this context, vehicles engage in a variety of interactions, some competitive, like overtaking and lane-changing, where each vehicle's aim may contradict another's, leading to a complex interplay as each strives for an advantageous position. On the flip side, there are cooperation interactions, which include platooning and collaborative driving, as well as the orchestrated dance of merging and diverging traffic. These interactions represent a symphony of shared goals, where vehicles work in concert to achieve greater efficiency, reduced energy consumption, or increased safety.

The traditional pair-wise approach falls short in capturing the full spectrum of these interactions. This is where hyperedges come into play. As a potent representation within hypergraphs, hyperedges can encapsulate these complex group-wise interactions by allowing for multiple nodes, or vehicles, to be connected simultaneously. This enables a more nuanced understanding of the collective behavior, revealing the underlying structure and dynamics of vehicular interactions. With

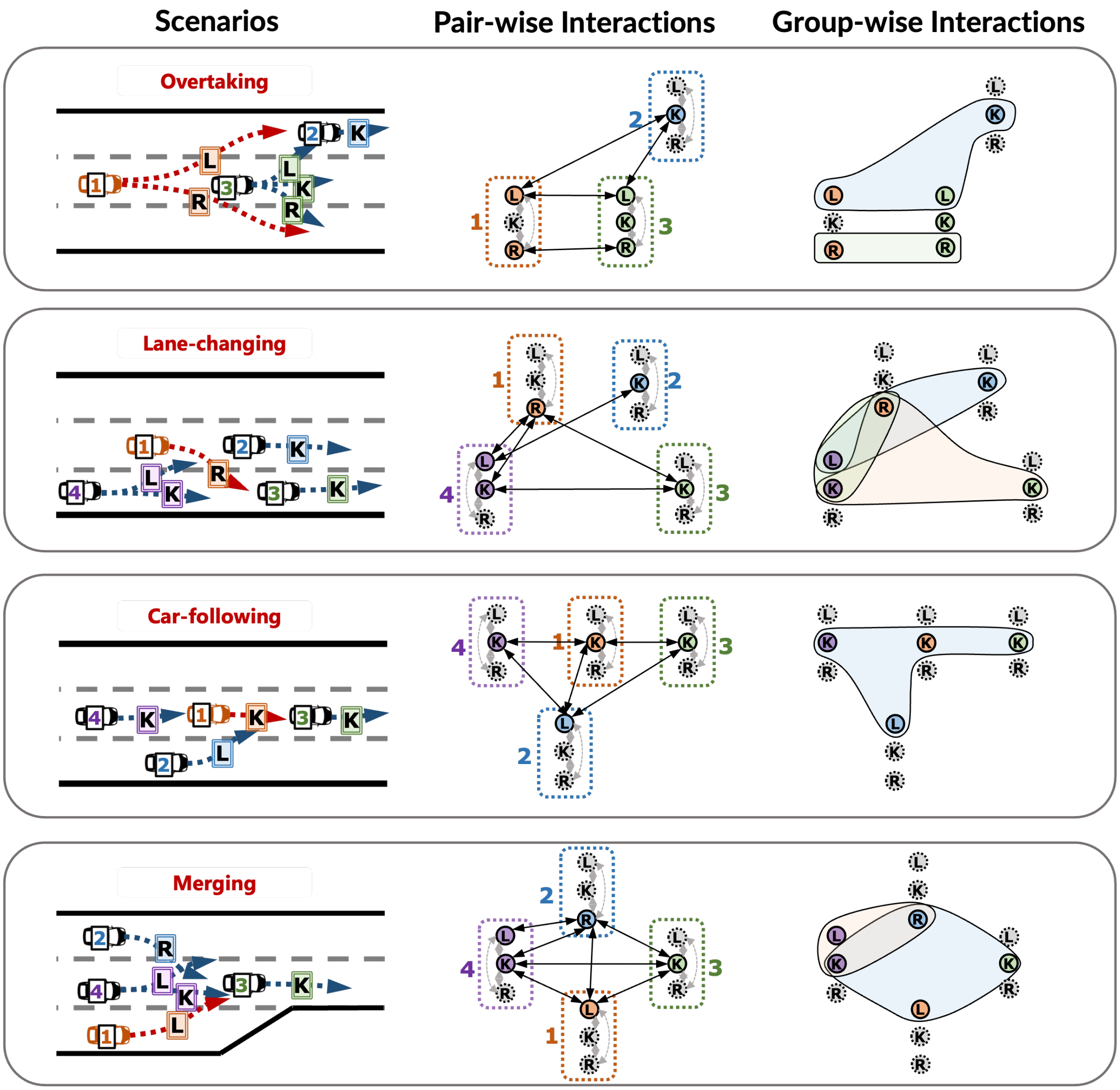


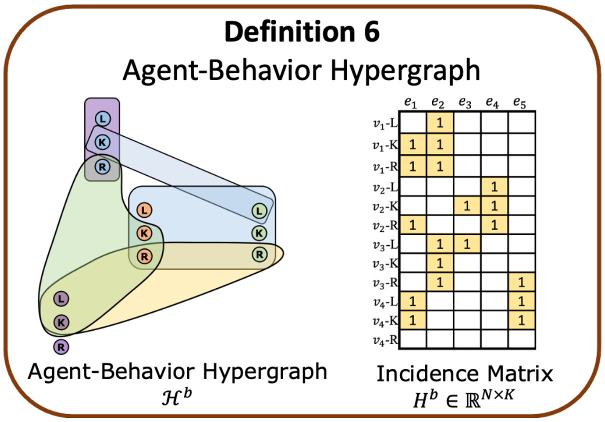
Figure 4.6: Insight 3: Pair-wise interactions and group-wise interactions among multiple possible behaviors of multiple vehicles in different multi-lane highway scenarios.

hyperedges, it is possible to model and analyze the rich patterns of interaction on highways, paving the way for more sophisticated traffic management systems that can optimize flow, enhance safety, and reduce congestion. Thus, we define an Agent-Behavior Hypergraph as the representation of the multi-agent multi-modal system for group-wise interaction relational reasoning.

Definition 6 (Agent-Behavior Hypergraph) *Let $\mathcal{H}^b$ be a hypergraph representation of the multi-modal motion states of N agents, with each of M behavior modes for each agent represented as a node. The hypergraph $\mathcal{H}^b$ is expressed as*

$$\mathcal{H}^b = (\mathcal{V}^b, \mathcal{E}^b; X^b, H^b)$$

where $\mathcal{V}^b \in \mathbb{R}^{|MN|}$ denotes the node set, $\mathcal{E}^b \in \mathbb{R}^K$ denotes the edge set (K can be complex), $X^b \in \mathbb{R}^{|MN| \times C}$ represents the feature tensor, and $H^b \in \mathbb{R}^{|MN| \times K}$ indicates the incidence matrix, where H^b_{ij} indicates whether node v_i is part of the hyperedge e_j.


Definition 6
Agent-Behavior Hypergraph
L
K
R
L
K
R
L
K
R
L
K
R
Agent-Behavior Hypergraph
$\mathcal{H}^b$
Incidence Matrix
$H^b \in \mathbb{R}^{N \times K}$

	e_1	e_2	e_3	e_4	e_5
v_1-L		1			
v_1-K	1	1			
v_1-R	1	1			
v_2-L				1	
v_2-K			1	1	
v_2-R	1			1	
v_3-L		1	1		
v_3-K		1			
v_3-R		1			1
v_4-L	1				1
v_4-K	1				1
v_4-R					

Figure 4.7: Agent-Behavior Hypergraph.

Figure 4.8 shows the relations between the six definitions of graphs and hypergraphs.

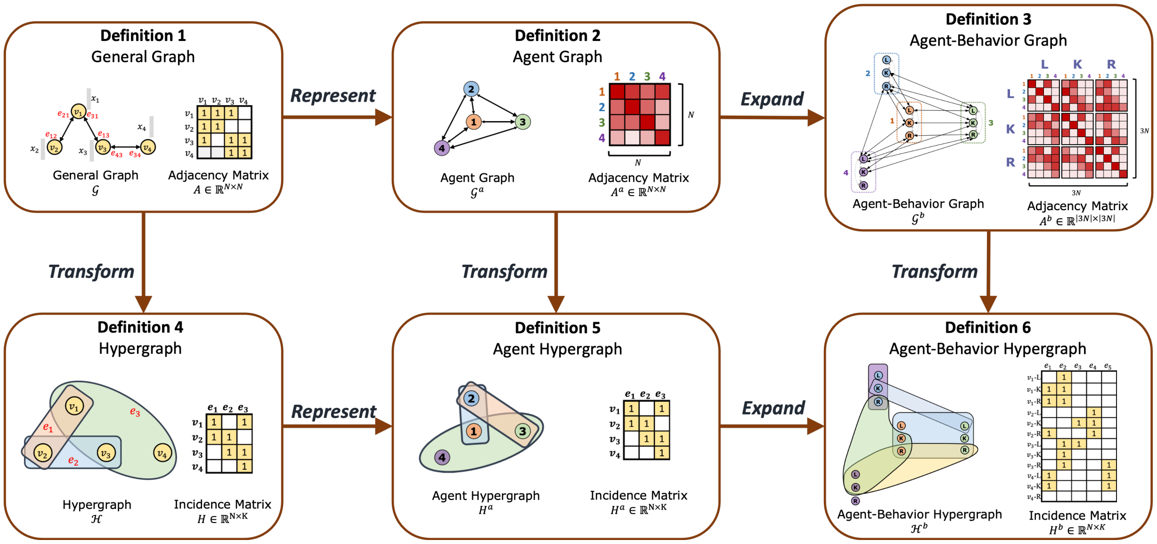


Figure 4.8: Relations between the definitions.

Chapter 5

Graph-based Interaction-aware Multi-agent Multi-modal Trajectory Prediction

5.1 GIMTP: Interaction-aware Multi-modal Trajectory Prediction

To achieve precise trajectory predictions in dense traffic scenarios, it is essential to capture the intricate temporal and social interactions between the target vehicle and its surrounding vehicles. To this end, we propose a `GIMTP` model composed of the following modules:

- **Dynamic Graph Embedding Module**: This module converts the original vehicle motion states into a dynamic graph embedding, taking into account the neighborhood, the distance, and the potential risk between vehicles.

- **Interaction Encoder**: This module captures the interactions between vehicles by employing a Diffusion Graph Convolution Network (DGCN) architecture, which is adept at handling the flow of information across the graph.
- **Intention Predictor**: This module maps the historical encodings and the guiding future trajectories to both lateral and longitudinal intentions over the future time horizon, effectively predicting the intended maneuvers.
- **Multi-modal Decoder**: This module merges the predicted intentions with the latent space, generating multiple future trajectory distributions with their corresponding probabilities.

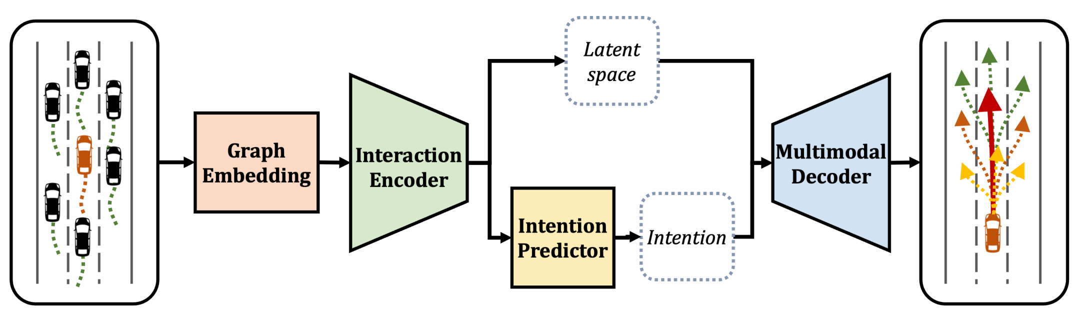


Figure 5.1: GIMTP framework architecture.

5.1.1 Dynamic Graph Embedding Module

Accurately predicting the trajectory of the target vehicle requires a deep understanding of its correlation and interaction with surrounding vehicles. As illustrated in Figure 5.2(a), surrounding vehicles can be classified longitudinally as preceding, parallel, and following vehicles, and laterally as left, same-lane, and right vehicles.

By stacking the motion states of the agent graphs along the temporal axis, we construct a dynamic spatial-temporal graph $\mathcal{G}_{1:T} = \{\mathcal{G}_t | \forall t \in 1, \ldots, T\}$ to represent the vehicle group's movement state from step 1 to step T. Specifically, the graph $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t; X_t, A_t)$ represents the motion states of the vehicle group at time step t. Each node v_t^i in the node set $\mathcal{V}_t = \{v_t^i | \forall i \in 1, \ldots, N\}$ represents each vehicle v^i in the vehicle group at step t. The edge set $\mathcal{E}_t$ indicates the influence between vehicles. A zero value for edge e^{ij} means there is no interaction between nodes v^i and v^j for $i, j \in 1, \ldots, N$. The feature matrix $X_t \in \mathbb{R}^{N \times C}$ contains the vehicle group's features, where C is the number of features.

Vehicle motion can significantly impact nearby vehicles. For example, a sudden lane change or speed alteration might force adjacent vehicles to slow down or change their path, while distant vehicles may remain unaffected. To quantitatively capture these interactions, we define a weighted adjacency matrix $A_t = \left(A_t^{ij}\right) \in \mathbb{R}^{N \times N}$ for the graph $\mathcal{G}_t$, where each element A_t^{ij} denotes the interaction intensity between vehicles v^i and v^j. This dynamic adjacency matrix is constructed by considering neighborhood, distance, and potential risks, accurately describing vehicle interactions. The following sections provide detailed insights into these factors.

Adjacency based on neighborhood: A vehicle's motion significantly impacts its immediate surroundings. To capture this, a binary adjacency matrix consider-

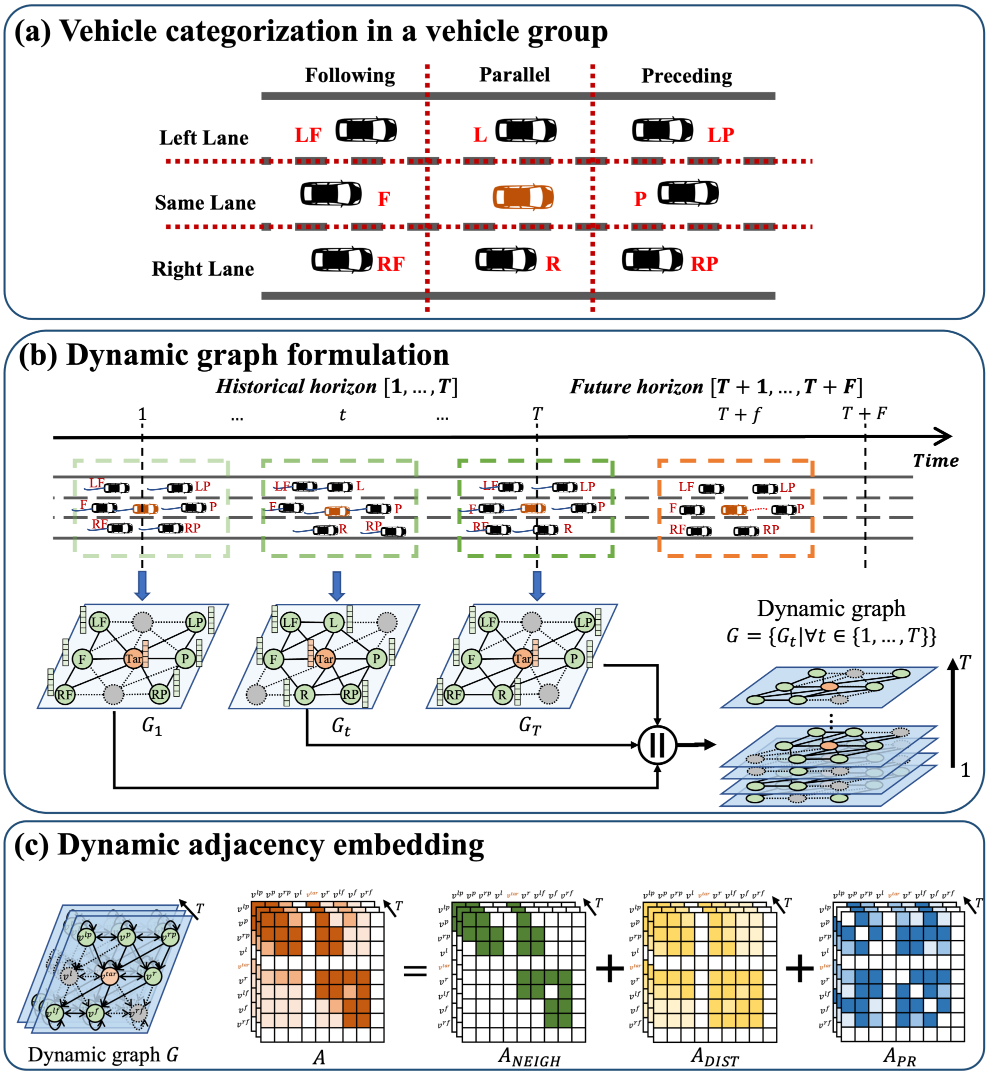

(a) Vehicle categorization in a vehicle group
Following
Parallel
Preceding
Left Lane
Same Lane
Right Lane
LF
L
LP
F
P
RF
R
RP
(b) Dynamic graph formulation
Historical horizon [1, …, T]
Future horizon [T + 1, …, T + F]
1
…
t
…
T
T + f
T + F
Time
Dynamic graph
$G = \{G_t | \forall t \in \{1, \dots, T\}\}$
G_1
G_t
G_T
Tar
T
1
(c) Dynamic adjacency embedding
Dynamic graph G
A
A_{NEIGH}
A_{DIST}
A_{PR}
=
+
+
T

Figure 5.2: Dynamic graph embedding of vehicle group's motion states.

ing immediate and preceding neighbors is defined as follows:

$$A_{t,NEIGH}^{ij} = \begin{cases} 1, & \text{if } v_t^i \text{ adjoins and precedes } v_t^j \\ 0, & \text{o.w.} \end{cases} \tag{5.1}$$

Adjacency dependent on distances: Intuitively, vehicles closer to each other exert a stronger influence on one another. We use a distance-decaying function to measure the weight between two vehicles, assigning higher weights to closer vehicles:

$$A_{t,DIST}^{ij} = \exp\left(-\left(\frac{dist(v_t^i, v_t^j)}{\sigma_{dist}}\right)^2\right) \tag{5.2}$$

where $dist(v_t^i, v_t^j)$ is the Euclidean distance between nodes v^i and v^j at step t, and σ_{dist} is the standard deviation of all distances between node pairs.

Adjacency dependent on potential risks: Potential fields are crucial for capturing social interactions [105, 106], extensively applied in autonomous vehicle path planning [107] and trajectory prediction [108]. We use the vehicles' kinetic energy to better represent risks and potential collisions. When a vehicle collides with another, its kinetic energy is transferred or converted, indicating an anomalous energy transfer process as per energy transfer theory [109]. Thus, potential traffic risk can be described as follows:

$$E^i = \frac{1}{2}m^i(s^i)^2 = \frac{1}{2}m^i s^i \cdot \frac{s^i - 0}{\Delta d^i} \cdot \Delta d^i \tag{5.3}$$

where E^i, m^i, s^i are the kinetic energy, mass, and velocity of vehicle v^i, respectively. Δd^i is the distance between the vehicle and another position in the traffic environment. Since $E^i = F^i \Delta d^i$, the equivalent force F^i caused by vehicle v^i can be represented as:

$$F^i = \frac{1}{2}m^i s^i \cdot \frac{s^i - 0}{\Delta d^i} \tag{5.4}$$

In a car-following scenario involving a follower vehicle v^i and a leader vehicle v^j, the traffic risk and the corresponding internal equivalent force between the two vehicles can be represented as:

$$E^{ij} = \frac{1}{2} m^i s^i \frac{s^i - s^j}{|d^i - d^j|} |d^i - d^j| \tag{5.5}$$

$$F^{ij} = \frac{1}{2} m^i s^i \frac{s^i - s^j}{|d^i - d^j|} \tag{5.6}$$

where s^j, d^j are the velocity and longitudinal position of vehicle v^j. The term $(s^i - s^j)/|d^i - d^j|$ indicates the relative velocity between vehicles v^i and v^j divided by their relative distance. The risk exists if the relative velocity is positive. This strategy can be extended to consider the risk in both longitudinal and lateral dimensions for the vehicle group. At each time step t, the longitudinal equivalent force and lateral equivalent force between two vehicles v^i and v^j can be represented as:

$$F^{ij}_{t,y} = \begin{cases} 0, & \text{if } s^i_{t,y} - s^j_{t,y} \leq 0 \\ \frac{1}{2} m^i s^i_{t,y} \frac{s^i_{t,y} - s^j_{t,y}}{|d^i_{t,y} - d^j_{t,y}|}, & \text{o.w.} \end{cases} \tag{5.7}$$

$$F^{ij}_{t,x} = \begin{cases} 0, & \text{if } s^i_{t,x} - s^j_{t,x} \leq 0 \\ \frac{1}{2} m^i s^i_{t,x} \frac{s^i_{t,x} - s^j_{t,x}}{|d^i_{t,x} - d^j_{t,x}|}, & \text{o.w.} \end{cases} \tag{5.8}$$

where $s^i_{t,x}, d^i_{t,x}$ denote the lateral speed and position of vehicle v^i at time step t, and $s^i_{t,y}, d^i_{t,y}$ denote its longitudinal velocity and position. The superscript represents the information of vehicle v^j at the same time step. The resultant force F^{ij} can be represented as:

$$F^{ij}_t = \sqrt{(F^{ij}_{t,x})^2 + (F^{ij}_{t,y})^2} \tag{5.9}$$

Thus, the adjacency matrix incorporating the potential risk can be represented as

$$A_{t,PR}^{ij} = \tanh\left(\frac{F_t^{ij}}{\sigma_F}\right) \tag{5.10}$$

where $\tanh(\cdot)$ is a hyperbolic tangent function, and σ_F is the standard deviation of all the forces between every pair of nodes.

Dynamic adjacency embedding: Let A be a dynamic adjacency matrix of dimension $\mathbb{R}^{T\times N\times N}$ which is formulated by concatenating $A_1, \ldots, A_T$. The matrix A represents a normalized aggregation of three adjacency matrices A_{NEIGH}, A_{DIST}, and A_{PR}. This aggregation is represented as:

$$\begin{aligned} A &= \text{normalize}(A_s) \\ &= \text{normalize}(A_{NEIGH} + A_{DIST} + A_{PR}) \end{aligned} \tag{5.11}$$

where the function normalize$(\cdot)$ is a min-max scaling normalization method which is applied to ensure that the elements of A reside within the bounded interval $[0, 1]$. A_s is the summation of the three adjacency matrices. Specifically, the normalization formula is given by

$$\text{normalize}(A_s) = \frac{A_s - \min(A_s)}{\max(A_s) - \min(A_s)} \tag{5.12}$$

with $\min(A_s)$ and $\max(A_s)$ representing the minimum and maximum values of all elements in A_s. Such a formulation of the dynamic adjacency matrix A provides a holistic and comprehensive representation of the dynamic interaction among vehicles.

It is important to note that the graph $\mathcal{G}$ is sparse and time-varying over the time horizon $[1, T]$. Since no single vehicle is assigned to any relative position in the vehicle group over time, vehicles do not always occupy each of the eight positions, resulting in graph sparsity. Furthermore, as vehicles move and distances between them change over time, the adjacency matrix continually varies. Surrounding vehicles may change their relative positions to the target vehicle. For example, a vehicle on the left of the target vehicle may accelerate and change lanes, moving from node v^l to node v^p. Vehicles can also leave the group, and new vehicles may join. This sparsity and variation over time characterize the microscopic interactions between vehicles.

5.1.2 Interaction Encoder

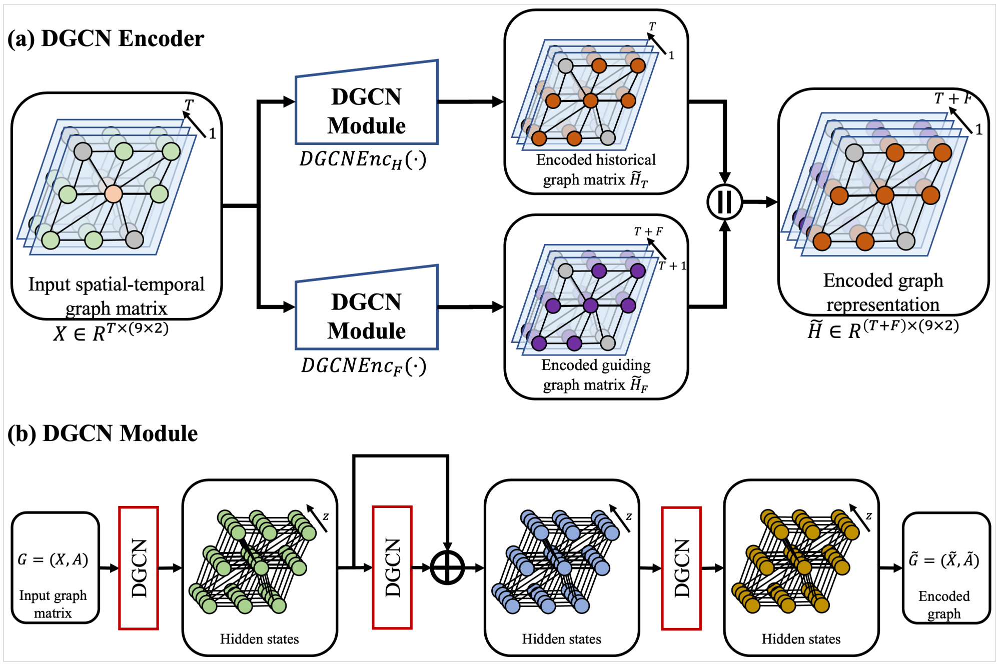

(a) DGCN Encoder
T
1
Input spatial-temporal graph matrix
$X \in R^{T\times(9\times2)}$
DGCN Module
$DGCNEnc_H(\cdot)$
DGCN Module
$DGCNEnc_F(\cdot)$
T
1
Encoded historical graph matrix $\tilde{H}_T$
$T+F$
$T+1$
Encoded guiding graph matrix $\tilde{H}_F$
$\|$
$T+F$
1
Encoded graph representation
$\tilde{H} \in R^{(T+F)\times(9\times2)}$
(b) DGCN Module
$G = (X, A)$
Input graph matrix
DGCN
Hidden states
z
DGCN
Hidden states
z
DGCN
Hidden states
z
$\tilde{G} = (\tilde{X}, \tilde{A})$
Encoded graph

Figure 5.3: Intention Encoder.

To effectively capture the bidirectional dependencies among nodes in the graph embedding, we incorporate the Diffusion Graph Convolutional Networks (DGCN) module, drawing inspiration from [110]. The DGCN layer is denoted as $DGCN^L(\cdot)$. The diffusion convolution is applied to the graph signal, encompassing both the forward diffusion process and its reverse:

$$\begin{aligned} H_{l+1} &= DGCN^L(H_l) \\ &= \sum_{k=1}^{K} \left(T_k(\bar{A}_f) \cdot H_l \cdot \Theta_{f,l}^k + T_k(\bar{A}_b) \cdot H_l \cdot \Theta_{b,l}^k \right) \end{aligned} \tag{5.13}$$

where the output from the l-th layer is represented by H_{l+1}. The masked feature matrix X serves as the input to the initial layer. The transformation block translating H_l to H_{l+1} is labeled as $DGCN(\cdot)$. The forward transition matrix, $\bar{A}_f = A/\text{rowsum}(A)$, captures dependencies from downstream nodes, while the backward transition matrix $\bar{A}_b = A^T/\text{rowsum}(A^T)$ reflects dependencies from upstream nodes. The function $T_k(\cdot)$ is a Chebyshev polynomial of order k, approximating the convolution operation involving the k-th layer neighbors of each node, expressed as $T_k(X) = 2X \cdot T_{k-1}(X) - T_{k-2}(X)$. $\Theta_{b,l}^k$ and $\Theta_{f,l}^k$ are the learnable parameters of the l-th layer, assigning weights to input data. For each vehicle in the vehicle group, the forward diffusion process captures influences from surrounding vehicles, while the reverse process captures influences it exerts on surrounding vehicles. As illustrated in Figure 5.4, this bidirectional diffusion convolution process in a DGCN layer incorporates influences from both upstream and downstream traffic flows [110, 111], capturing inherent bidirectional and heterogeneous interactions between the target vehicle and its surrounding vehicles.

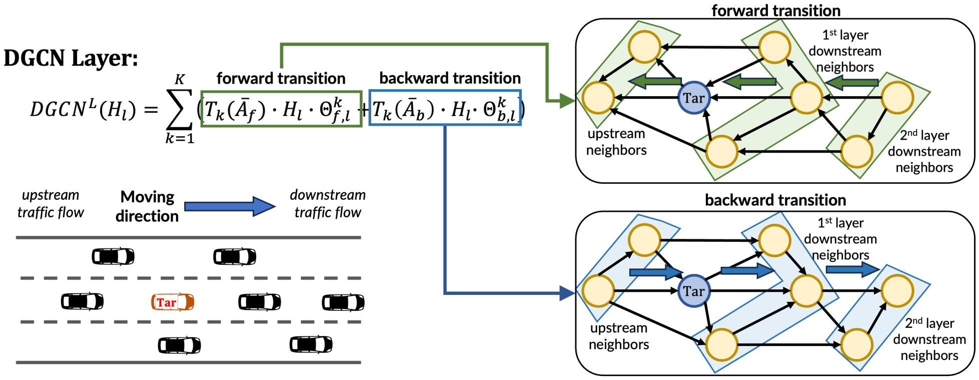

DGCN Layer:
$DGCN^L(H_l) = \sum_{k=1}^{K} (T_k(\bar{A}_f) \cdot H_l \cdot \Theta_{f,l}^k + T_k(\bar{A}_b) \cdot H_l \cdot \Theta_{b,l}^k)$
forward transition
backward transition
upstream traffic flow
Moving direction
downstream traffic flow
Tar
forward transition
1st layer downstream neighbors
upstream neighbors
2nd layer downstream neighbors
backward transition
1st layer downstream neighbors
upstream neighbors
2nd layer downstream neighbors

Figure 5.4: Forward and backward diffusion convolution process in a DGCN layer.

The calculation process of the three-layer DGCN module in the encoder can be summarized as follows:

$$H_1 = DGCN_1^L(X) \tag{5.14}$$

$$H_2 = \sigma(DGCN_2^L(H_1)) + H_1 \tag{5.15}$$

$$H_o = DGCN_3^L(H_2) \tag{5.16}$$

where H_1 and H_2 are the outputs of the first and second DGCN layers, respectively. The hidden state H_o is the encoded feature matrix of the graph $\tilde{G}$, which is the output of the DGCN module.

We employ two DGCN modules, each containing a three-layer DGCN architecture, to capture the bidirectional propagation among the vehicle group. One module encodes the historical states of the vehicle group in the historical time horizon $[1, \ldots, T]$ into a historical graph embedding $\tilde{H}_T$, while the other maps the vehicle group's historical states to the future time horizon $[T+1, \ldots, T+F]$ as a future-guided graph embedding $\tilde{H}_F$.

$$\tilde{H}_T = DGCNEnc_H(X) \tag{5.17}$$

$$\tilde{H}_F = DGCNEnc_F(X) \tag{5.18}$$

$$\tilde{H} = [\tilde{H}_T, \tilde{H}_F] \tag{5.19}$$

The resulting graph matrices from both modules are merged to form an integrated embedding $\tilde{H}$, covering a time window $[1, T+F]$ that spans both historical and future horizons.

5.1.3 Intention Predictor

Although vehicle motion states can be ambiguous due to the multifaceted and manifold potential driving intentions in multi-lane highway environments, the spectrum of driving intentions is rather finite. Therefore, we categorize them as follows: laterally, into lane keeping (LK), left lane change (LLC), and right lane change (RLC), denoted as $m^{lat} = [m^{LK}, m^{LLC}, m^{RLC}] \in \mathbb{R}^{3\times F}$; longitudinally, into constant speed (CS), acceleration (ACC), and deceleration (DEC), denoted as $m^{lon} = [m^{CS}, m^{ACC}, m^{DEC}] \in \mathbb{R}^{3\times F}$. Denote the ground-truth intention set as $M = [m^{lat}, m^{lon}] \in \mathbb{R}^{6\times F}$. We employ a one-hot encoding strategy for both m^{lat} and m^{lon} at every time step in future time horizon F.

We treat the intention prediction as a classification problem over the future horizon $[T+1, T+F]$. Our intention classification approach draws inspiration and parallels the methodology detailed in [19]. As presented in Figure 5.5, we adopt the encoded graph representation $\tilde{H}$ as the input. We first incorporate two

MLP layers to reduce the dimensionality and further encode the tensor to a latent space with hidden size z. To extract the probability distributions related to lateral and longitudinal intention classes, we adopt a LatMLP and a LonMLP layer with softmax activation function to respectively generate lateral and longitudinal intention classification along the time axis F. The mathematical formulations are:

$$H_2^{IP} = MLP_2^{IP}(MLP_1^{IP}(\tilde{H})) \tag{5.20}$$

$$\hat{m}^{lat} = \text{softmax}(LatMLP(H_2^{IP})) \tag{5.21}$$

$$\hat{m}^{lon} = \text{softmax}(LonMLP(H_2^{IP})) \tag{5.22}$$

where MLP_1^{IP} aggregates the hidden states of the vehicle groups and MLP_2^{IP} maps the input time horizon $[1, T+F]$ to the output time horizon $[T+1, T+F]$. $\hat{m}^{lat}$ and $\hat{m}^{lon}$ denote the predicted lateral and longitudinal intentions, respectively.

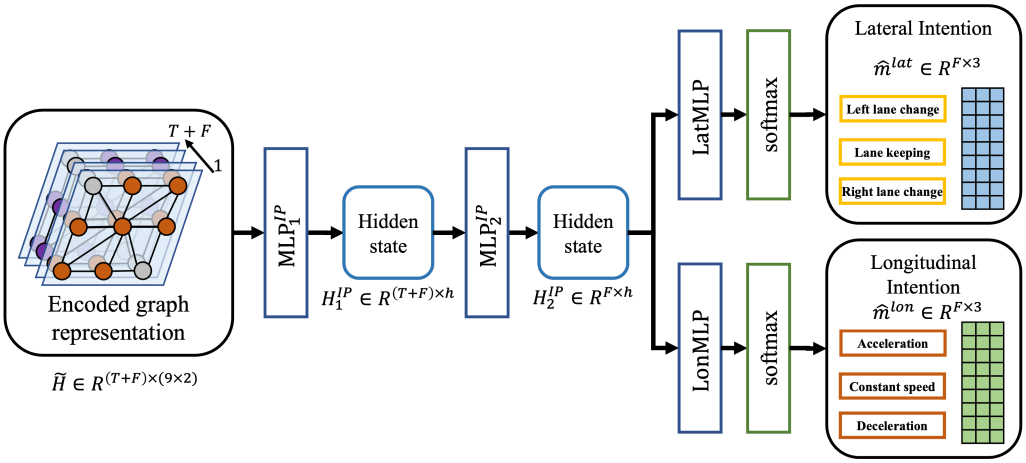


Figure 5.5: Intention Predictor.

5.1.4 Multi-modal Decoder

Feature vectors derived from the encoder over different historical time steps within $[1, T]$ demonstrate varying impacts on motion states at distinct steps in the future horizon $[T + 1, T + F]$. Given the inherent sequential nature of motion, a vehicle's movement patterns at the time mark $T + 1$ exhibit a more pronounced association with its immediate preceding time steps adjacent to T compared to those at temporally distant steps. Nonetheless, the predicted trajectories modulated by diverse intentions exhibit heterogeneity, emphasizing the importance of feature fusion. To capture the interrelation between each time step in the input horizon, which spans both historical and future horizons, and each time step in the output future horizon, we use an enhanced adaptation of the intention-specific feature fusion proposed by [50]. This method explicitly considers the relevance of the encoded features by fusing the feature vectors from both distinct historical and future time steps for each intention, fostering an enriched comprehension of their impact on future states.

As presented in Figure 5.6, for each predicted longitudinal and lateral intention vector, denoted as $m \in \mathbb{R}^F$, within the intention matrix $\hat{M} = [\hat{m}^{lon}, \hat{m}^{lat}] \in \mathbb{R}^{F \times 6}$, a corresponding trainable weight matrix $W_m \in \mathbb{R}^{(T+F) \times F}$ is established. The matrix $W_m = \left[u^m_{t,t'} | t \in [1, T + F], t' \in [T + 1, T + F]\right]$ for each intention consists of weight $u_{t,t'}$ that apportions the influence of the state at a given t to the state at t'. Specifically, the weight matrix W_m corresponding to each intention can be articulated and structured as follows in the matrix representation:

$$W_m = \begin{bmatrix} u^m_{1,1} & \dots & u^m_{1,F} \\ \vdots & \dots & \vdots \\ u^m_{T,1} & \dots & u^m_{T,F} \\ u^m_{T+1,1} & \dots & u^m_{T+1,F} \\ \vdots & \dots & \vdots \\ u^m_{T+F,1} & \dots & u^m_{T+F,F} \end{bmatrix} \in \mathbb{R}^{(T+F)\times F} \tag{5.23}$$

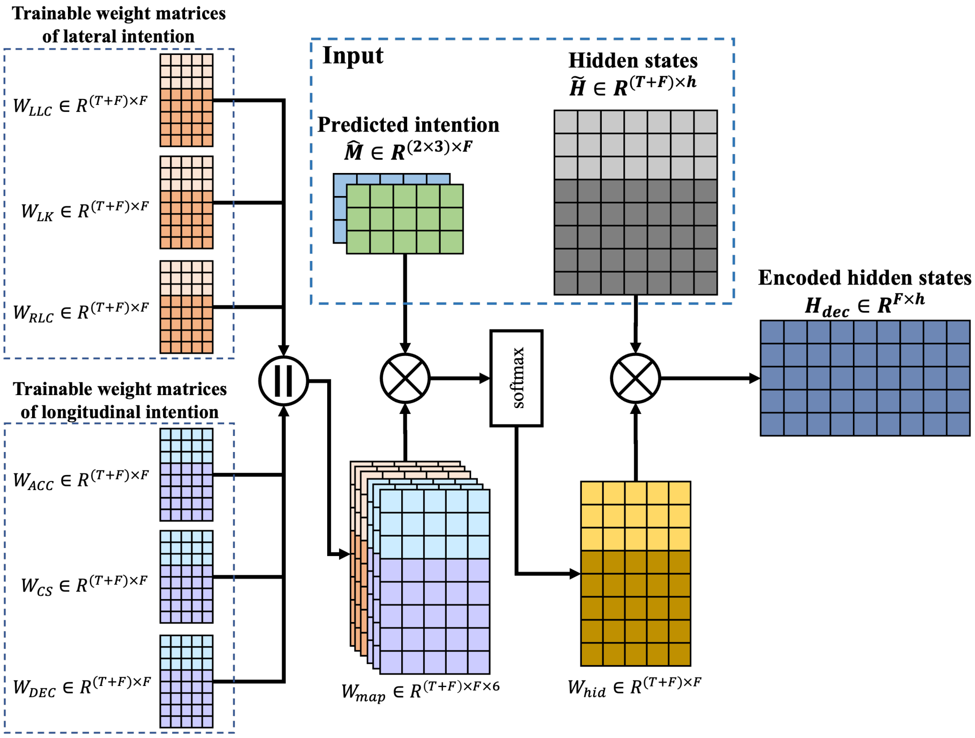

Trainable weight matrices of lateral intention
$W_{LLC} \in R^{(T+F)\times F}$
$W_{LK} \in R^{(T+F)\times F}$
$W_{RLC} \in R^{(T+F)\times F}$
Trainable weight matrices of longitudinal intention
$W_{ACC} \in R^{(T+F)\times F}$
$W_{CS} \in R^{(T+F)\times F}$
$W_{DEC} \in R^{(T+F)\times F}$
Input
Predicted intention
$\widehat{M} \in R^{(2\times 3)\times F}$
Hidden states
$\widetilde{H} \in R^{(T+F)\times h}$
softmax
$W_{map} \in R^{(T+F)\times F\times 6}$
$W_{hid} \in R^{(T+F)\times F}$
Encoded hidden states
$H_{dec} \in R^{F\times h}$

Figure 5.6: Intention feature fusion.

TThe weight matrices of the six intentions are stacked as $W_{map} \in \mathbb{R}^{(T+F)\times F\times 6}$. A batch matrix multiplication $\otimes$ paired with softmax activation seamlessly emphasizes the six intentions' contributions:

$$W_{hid} = \text{softmax}(W_{map} \otimes \hat{M}) \tag{5.24}$$

Given the encoded feature vectors $\tilde{H} \in \mathbb{R}^{(T+F)\times z}$ acquired from the Interaction Encoder, the weight matrix $W_{hid} \in \mathbb{R}^{(T+F)\times F}$ is multiplied to harmonize the influence of the intentions with latent states, subsequently mapping this confluence to the states within the predictive horizon $[T+1,\ldots,T+F]$. This fusion results in the generation of the encoded latent states H_{dec} for the Multi-modal Decoder:

$$H_{dec} = W_{hid} \cdot \tilde{H} \tag{5.25}$$

To approximate the probabilistic trajectory distribution, we adhere to the foundational principles of the total probability theorem and decompose $P(Y|X)$ in the following manner:

$$P(Y|X) = P_{\theta}(Y|X, M)P(M|X) \tag{5.26}$$

The output $\{Y_{t'}|\forall t' \in [T+1,\ldots,T+F]\}$ consists of a five dimentional vector governing a bivariate Gaussian distribution of the lateral and longitudinal position: mean $\mu_{t',x}, \mu_{t',y}$, standard deviation $\sigma^2_{t',x}, \sigma^2_{t',y}$, and correlation coefficient $\rho_{t'}$.

As presented in Figure 5.7, the fused feature H_{dec} coupled with the probabilistically inferred intention matrix $\hat{M}$ are concatenated and then fed into an MLP layer MLP_1^{MD}. The encoded feature states H_1^{MD} are successively channeled into a decoder structured on the GRU architecture considering the temporal continuity of predicted trajectory. We further adopt MLP_2^{MD} to transform the hidden state H_2^{MD} into the five parameters of the bivariate Gaussian distribution. In this context:

$$H_1^{MD} = MLP_1^{MD}(H_{dec}, \hat{M}) \tag{5.27}$$

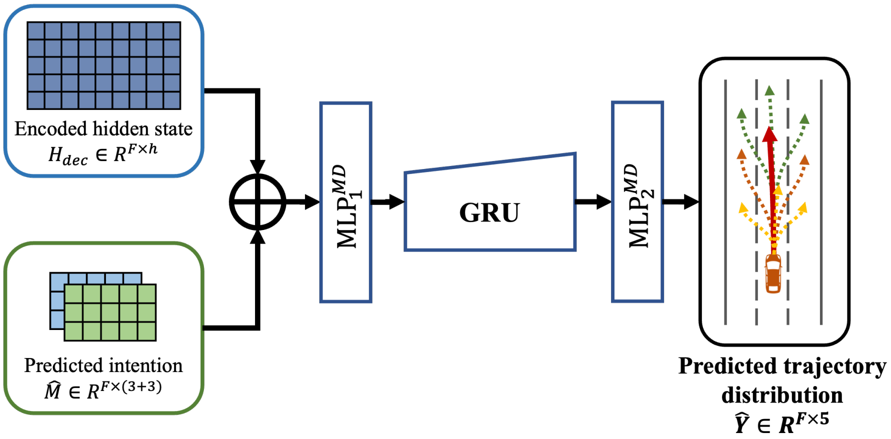

Encoded hidden state
$H_{dec} \in R^{F \times h}$
Predicted intention
$\widehat{M} \in R^{F \times (3+3)}$
MLP_1^{MD}
GRU
MLP_2^{MD}
Predicted trajectory distribution
$\widehat{Y} \in R^{F \times 5}$

Figure 5.7: Multi-modal Decoder.

$$H_2^{MD} = GRU(H_1^{MD}) \tag{5.28}$$

$$\hat{Y} = MLP_2^{MD}(H_2^{MD}) \tag{5.29}$$

where $\hat{Y}$ denotes the final output yielded by the model, signifying the bivariate Gaussian distribution associated with the predicted position.

5.2 GIRAFFE: Interaction-aware Multi-agent Multi-modal Trajectory Prediction

As the first component of the framework, accurate multi-modal trajectory prediction based on the probabilistic modeling of various behaviors for multiple agents is essential. To achieve this, we adopted an enhanced version of `GIRAFFE`, building upon the `GIMTP` framework proposed in [29]. `GIRAFFE` is specifically designed to capture the interactions between the target vehicle and its surrounding vehicles.

We improved the multi-agent prediction capabilities and optimized computational efficiency by pruning layers through a combination of the following modules:

- **Interaction Encoder**: This module captures the interactions between vehicles by employing a light-weight Diffusion Graph Convolution Network (DGCN) architecture, which is adept at handling the flow of information across the graph.
- **Intention Predictor**: This module maps the historical encodings and the guiding future trajectories to provide only lateral intentions over the future time horizon, effectively predicting the intentions.
- **Multi-modal Decoder**: This module merges the predicted intentions with the latent space, generating multiple future trajectory distributions with their corresponding probabilities for all the vehicles in the vehicle group.

The process begins with constructing a dynamic spatial-temporal graph as the model input, where each node represents a vehicle, and edges denote the interactions between vehicles. These interactions are weighted by factors such as neighborhood proximity, distance, and potential collision risks. The Interaction Encoder, utilizing a DGCN architecture, encodes these dynamic graph embeddings. The DGCN captures bidirectional dependencies among vehicles by applying diffusion convolutions, which consider both forward and reverse processes to model the influence of surrounding vehicles and the target vehicle's impact on them. This encoder generates graph embeddings for both historical states and future-guided

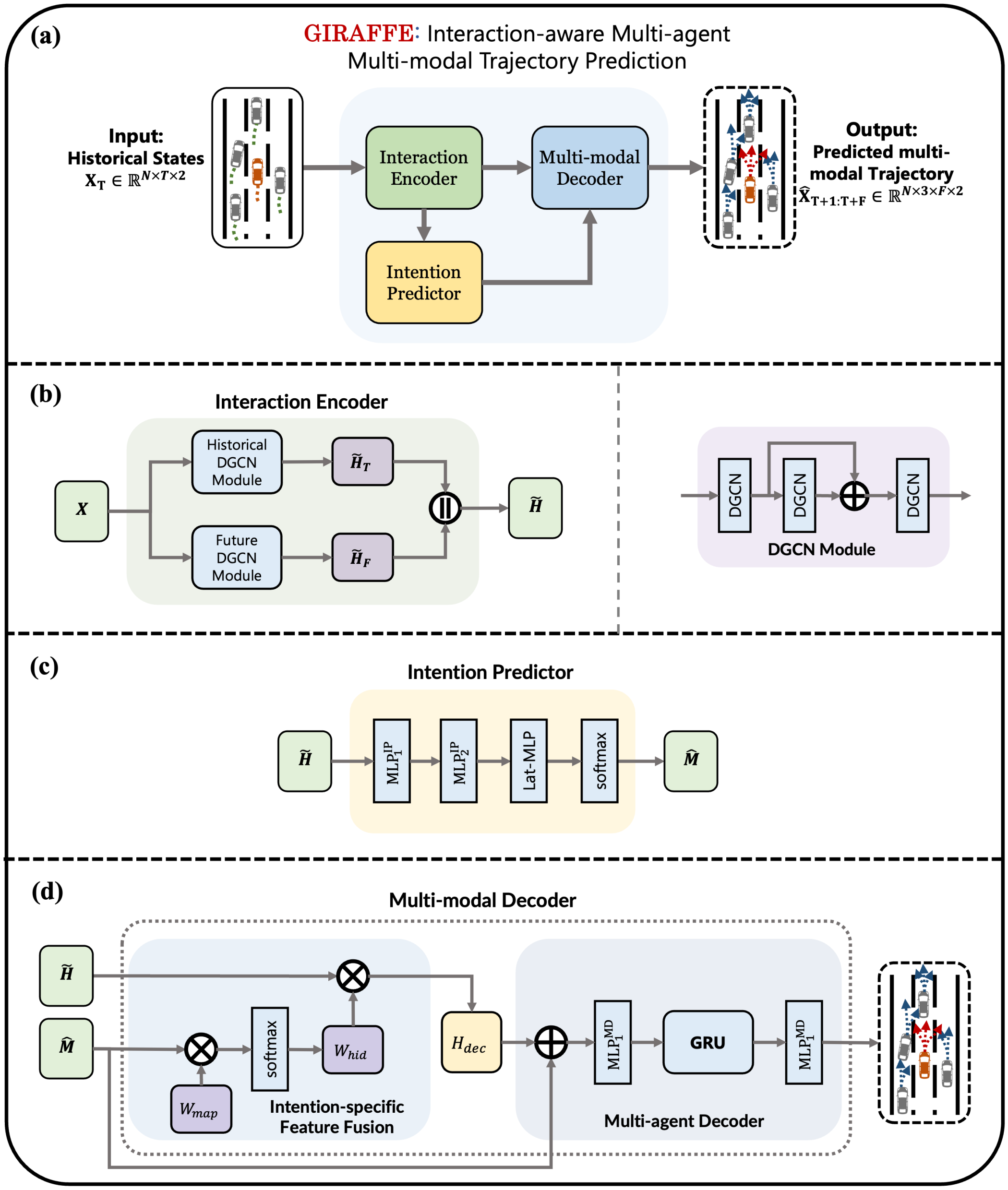


Figure 5.8: GIRAFFE framework architecture.

trajectories, merging them into a comprehensive representation that spans the entire time window of interest.

The Intention Predictor tackles the classification of future driving intentions, both laterally and longitudinally. Using the encoded graph representation, two MLP layers reduce the dimensions and encode the features into a latent space. The LatMLP and LonMLP layers, equipped with softmax activation, then classify the lateral and longitudinal intentions over the future time horizon. These predictions are crucial for understanding potential maneuvers the vehicle might take, such as lane changes or speed adjustments.

Finally, the Multi-modal Decoder combines the predicted intentions of multiple agents with the latent space to produce multiple future trajectory distributions for each agent. This decoder employs a trainable weight matrix to merge features from distinct historical and future time steps, emphasizing the importance of sequential motion patterns. The GRU-based decoder ensures temporal continuity in the predicted trajectories, mapping the fused features to a bivariate Gaussian distribution representing the future vehicle positions. This approach enables the model to generate probabilistic predictions for multiple agents, enhancing the overall accuracy and reliability of trajectory forecasting.

Chapter 6

Hypergraph-based Interaction Relational Reasoning Motion Generation and Planning

The core of RHINO is the learning of a multiscale hypergraph, where nodes represent agent behaviors and hyperedges capture their group-wise interactions. This hypergraph is utilized to derive agent and interaction embeddings, thereby providing a deeper understanding of the underlying relational dynamics. Additionally, we incorporate a multi-agent trajectory generation system based on the CVAE framework to manage the stochasticity of each agent's potential behaviors and motion states, producing plausible trajectories for each vehicle. To ensure precise trajectory planning in dense traffic for the target vehicle, a neural planner evaluates and selects the optimal trajectory, refining it for efficiency, comfort, and safety.

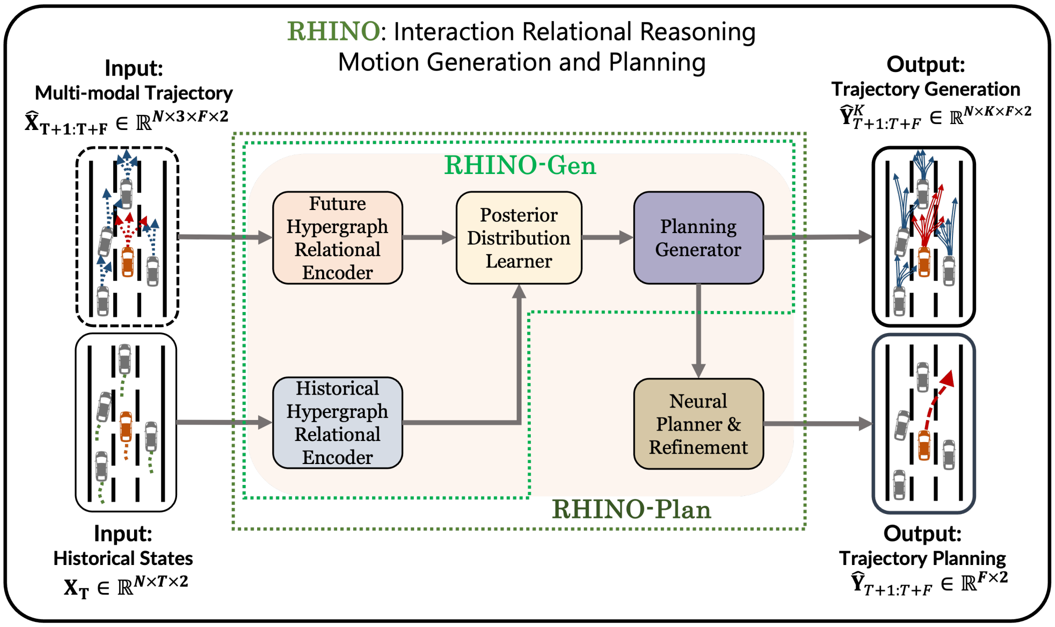

RHINO: Interaction Relational Reasoning
Motion Generation and Planning
Input:
Multi-modal Trajectory
$\hat{\mathbf{X}}_{\mathbf{T+1:T+F}} \in \mathbb{R}^{N\times 3\times F\times 2}$
Output:
Trajectory Generation
$\hat{\mathbf{Y}}^{K}_{T+1:T+F} \in \mathbb{R}^{N\times K\times F\times 2}$
RHINO-Gen
Future Hypergraph Relational Encoder
Posterior Distribution Learner
Planning Generator
Historical Hypergraph Relational Encoder
Neural Planner & Refinement
RHINO-Plan
Input:
Historical States
$\mathbf{X}_{\mathbf{T}} \in \mathbb{R}^{N\times T\times 2}$
Output:
Trajectory Planning
$\hat{\mathbf{Y}}_{T+1:T+F} \in \mathbb{R}^{F\times 2}$

Figure 6.1: RHINO Framework.

Thus, `RHINO` comprises the following modules:

- **Hypergraph Relational Encoder**, which transforms both the original historical states and predicted multi-agent multi-modal trajectories into hypergraphs, modeling and reasoning the underlying relations between the vehicles.

- **Posterior Distribution Learner**, which captures the posterior distribution of the future trajectory given the historical states and the predicted multi-modal future motion states of all the vehicles in the vehicle group.

- **Residual Decoder**, which decodes the embeddings by concurrently reconstructing the historical states and generating the future trajectories.

- **Neural Planner and Refinement**, which selects and optimizes the generated trajectory planning based on all the generated possible trajectories of all the vehicles considering safety, efficiency, and comfort.

To clarify and distinguish between motion generation for all vehicles in the group and motion planning for the target vehicle, we developed two versions of the model: `RHINO-Gen` and `RHINO-Plan`. `RHINO-Gen` includes the Hypergraph Relational Encoder, Posterior Distribution Learner, and Residual Decoder modules. `RHINO-Plan` includes all these modules plus the Neural Planner and Refinement module.

6.1 Hypergraph Relational Encoder

We employ two Hypergraph Relational Encoder modules: a Historical Hypergraph Relational Encoder for handling historical states and a Future Hypergraph Relational Encoder for predicted multi-agent multi-modal trajectories from `GIRAFFE`. For the Historical Hypergraph Relational Encoder, the input historical states X_T form an Agent Hypergraph $\mathcal{H}_T^a$. For the Future Hypergraph Relational Encoder, the predicted multi-agent multi-modal trajectories $\hat{X}_{T+1:T+F}$ form an Agent-Behavior Hypergraph $\mathcal{H}_T^b$, where each agent node is expanded into three lateral behavior nodes. Both modules share the same structure regardless of the input hypergraph types.

6.1.1 Multiscale Hypergraph Topology Inference

To capture group-wise interactions in hypergraphs across various scales, we infer a multiscale hypergraph that reflects interactions at different group sizes. Let $\mathcal{H} = \{\mathcal{H}^{(0)}, \mathcal{H}^{(1)}, \cdots, \mathcal{H}^{(S)}\}$ be a multiscale hypergraph, and $\mathcal{V} = \{v_1, v_2, \cdots, v_N\}$ be a set of nodes. At any scale s, $\mathcal{H}^{(s)} = (\mathcal{V}, \mathcal{E}^{(s)})$ has a hyperedge set $\mathcal{E}^{(s)} = \{e_1^{(s)}, e_2^{(s)}, \cdots, e_K^{(s)}\}$ representing group-wise relations with K hyperedges. A larger s indicates a larger scale of agent groups, while $\mathcal{H}^{(0)} = (\mathcal{V}, \mathcal{E}^{(0)})$ models the finest pair-wise agent connections. The topology of each $\mathcal{H}^{(s)}$ is represented as an incidence matrix $H^{(s)}$.

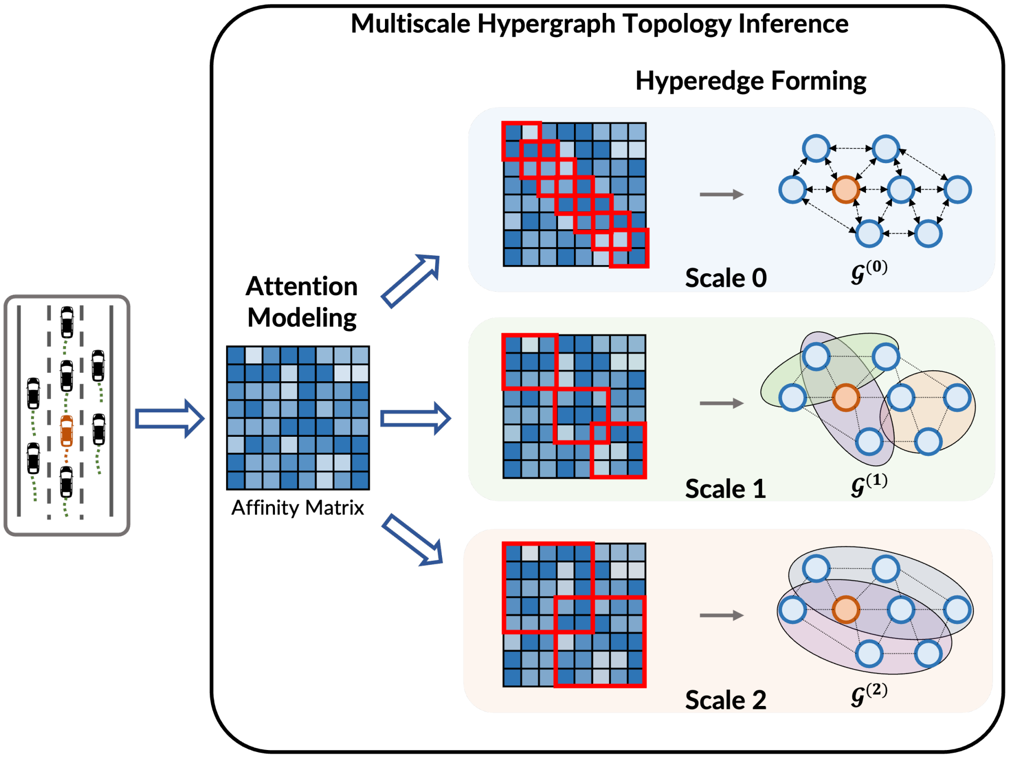


Figure 6.2: Hypergraph encoder.

6.1.1.1 Affinicy Modeling.

To understand and quantify dynamic interactions between agents, we employ trajectory embedding to distill motion states into compact, informative representations. To infer a multiscale hypergraph, we construct hyperedges by grouping agents with highly correlated trajectories, represented as high-dimensional feature vectors. For the i-th agent, the trajectory embedding is denoted as q_i, a function of the agent's state history over a temporal window from time 1 to time T. The embedding function f_Q, an MLP, transforms the motion states X^i into a vector in $\mathbb{R}^d$. Mathematically, this is represented as:

$$q_i = f_Q(X^i) \in \mathbb{R}^d \tag{6.1}$$

The affinity between agents is represented by an affinity matrix A, containing the pairwise relational weights between all agents. The affinity matrix is defined as:

$$A = \{A_{ij} | i, j = 1, \ldots, N\} \in \mathbb{R}^{N \times N} \tag{6.2}$$

Each element A_{ij} is computed as the correlation between the trajectory embeddings of the i-th and j-th agents. The correlation is the normalized dot product of the two trajectory embeddings, expressed as:

$$A_{ij} = \frac{q_i^\top q_j}{\|q_i\|_2 \|q_j\|_2} \tag{6.3}$$

Here, $\|\cdot\|_2$ denotes the L2 norm. The relational weight A_{ij} measures the strength of association between the trajectories of the i-th and j-th agents, capturing the degree to which their behaviors are correlated. This enables the assessment of interaction patterns and can uncover underlying social or physical laws governing agent dynamics.

6.1.1.2 Hyperedge Forming.

Formulating a hypergraph necessitates the strategic formation of hyperedges that reflect the complex interactions between the nodes in the system. Initially, the 0-th scale hypergraph $\mathcal{H}^{(0)}$ is considered, where the construction is based on pair-wise connections. Each node establishes a link with another node that has the highest affinity score with it.

As the system's complexity scales up, starting from scale $s \geq 1$, the methodology shifts towards group-wise connections. This shift is based on the intuition that agents within a particular group should display strong mutual correlations, indicating a propensity for concerted action. To implement this, a sequence of increasing group sizes $\{K^{(s)}\}_{s=1}^{S}$ is established. For every node, denoted by v_i, the objective is to discern a group of agents that are highly correlated, ultimately forming $K^{(s)}$ groups or hyperedges at each scale s. The hyperedge associated with a node v_i at a given scale s is indicated by $e_i^{(s)}$. The determination of the most correlated agents is framed as an optimization problem, aiming to link these agents into a hyperedge that accounts for group dynamics:

$$e_i^{(s)} = \underset{\Omega \subseteq \mathcal{V}}{\text{arg max}} \, \|A_{\Omega,\Omega}\|_{1,1} \tag{6.4}$$

$$\text{s.t. } |\Omega| = K^{(s)}; v_i \in \Omega; i = 1, \ldots, N \tag{6.5}$$

The culmination of this hierarchical structuring is a multiscale hypergraph, encapsulated by the set $\{\mathcal{H}^{(s)} \in \mathbb{R}^{N \times N}\}_{s=1}^{S}$, where each scale s embodies a distinct layer of abstraction in representing agent relationships within the hypergraph.

Multiscale hyperedge formation is computationally efficient for identifying high-order relationships from a single matrix, and it ensures stable and informative training of the affinity matrix through back-propagation.

6.1.2 Hypergraph Neural Message Passing

To uncover patterns in agent motion states from the inferred multiscale hypergraph, we develop a multiscale hypergraph neural message passing technique. This iteratively computes agent and interaction embeddings through node-to-hyperedge and hyperedge-to-node processes, as shown in Figure 6.3. Initially, each agent's embedding is derived from its trajectory. For any given scale, the initial embedding for the i-th agent, v_i, is set as $v_i = q_i \in \mathbb{R}^d$. During the node-to-hyperedge phase, agent embeddings are aggregated to generate interaction embeddings. Subsequently, in the hyperedge-to-node phase, each agent's embedding is updated based on the associated interaction embeddings. This iterative process refines agent embeddings by considering evolving relationships encapsulated within hyperedges.

Node-to-Hyperedge Mapping nodes to hyperedges is crucial in constructing the hypergraph topology. Each node v_j is associated with a hyperedge e_i if v_j is an element of e_i. This mapping defines the hyperedge interaction embedding. The interaction embedding for a hyperedge e_i is a function of the embeddings of its con-

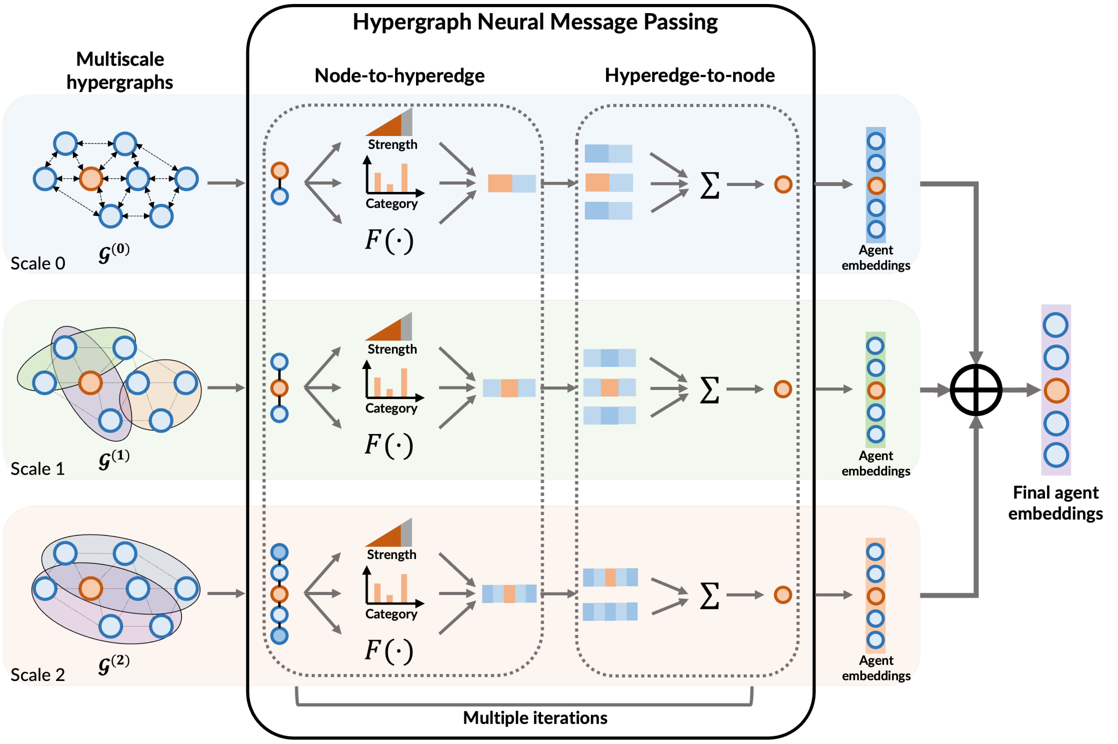


Figure 6.3: Hypergraph encoder.

stituent nodes, modulated by the neural interaction strength r_i and categorized by coefficients $c_{i,l}$. The per-category function $\mathcal{F}_l$ models each interaction category and processes the aggregated node embeddings. Each $\mathcal{F}_l$ is a trainable MLP, processing the node embeddings within a specific interaction category. The mathematical formulation is:

$$e_i = r_i \sum_{l=1}^{L} c_{i,l} \mathcal{F}_l \left(\sum_{v_j \in e_i} v_j \right) \in \mathbb{R}^d \tag{6.6}$$

The neural interaction strength r_i captures the intensity of interaction within the hyperedge and is obtained through a trainable model $\mathcal{F}_r$, applied to a collective embedding z_i with a sigmoid function σ. This collective embedding z_i is the weighted sum of the individual node embeddings within the hyperedge. The weight w_j for each node is determined by a trainable MLP $\mathcal{F}_w$. The mathematical formulations are:

$$r_i = \sigma(\mathcal{F}_r(z_i)) \tag{6.7}$$

$$z_i = \sum_{v_j \in e_i} w_j v_j \tag{6.8}$$

$$w_j = \mathcal{F}_w \left(v_j, \sum_{v_m \in e_i} v_m \right) \tag{6.9}$$

The neural interaction category coefficients $c_{i,l}$ denote the probability of the l-th neural interaction category within L possible categories. These coefficients are computed using a softmax function applied to the output of another trainable MLP $\mathcal{F}_c$, adjusted by a Gumbel distribution g and a temperature parameter τ:

$$c_i = \text{softmax} \left(\frac{\mathcal{F}_c(z_i) + g}{\tau} \right) \tag{6.10}$$

Hyperedge-to-Node The hyperedge-to-node mapping updates and refines agent embeddings within the hypergraph framework. Each hyperedge e_j maps back to its constituent nodes v_i, assuming v_i is included in e_j. The primary goal is to update the agent's embedding using the function $\mathcal{F}_v$, a trainable MLP. The updated agent embedding v_i results from applying $\mathcal{F}_v$ to the concatenation of the agent's current embedding and the sum of the embeddings of all hyperedges the agent is part of. Formally, the update rule is:

$$v_i \leftarrow \mathcal{F}_v\left(\left[v_i, \sum_{e_j \in \mathcal{E}_i} e_j\right]\right) \in \mathbb{R}^d \tag{6.11}$$

where $\mathcal{E}_i = \{e_j \mid v_i \in e_j\}$ denotes the set of hyperedges associated with the i-th node v_i, and the square brackets $[\cdot, \cdot]$ symbolize the operation of embedding concatenation. This operation merges the individual node embedding with the collective information conveyed by the associated hyperedges, encapsulating the influence exerted by the hyperedges on the individual agent.

The framework iteratively applies the node-to-hyperedge and hyperedge-to-node phases multiple times. This iterative process refines agent embeddings by considering the evolving relationships encapsulated within hyperedges.

Upon completion of these iterations, the output is constructed as the concatenation of the agent embeddings across all scales. The final agent embedding matrix $\mathbf{V}$ comprises the embeddings of all agents, where each agent embedding v_i is a concatenation of the embeddings from all scales. This is mathematically expressed as:

$$\mathbf{V} = [v_i], \quad \forall i \in [1, \ldots, N] \in \mathbb{R}^{N \times |d(S+1)|} \tag{6.12}$$

where

$$v_i = [v_i^{(0)}, v_i^{(1)}, \ldots, v_i^{(S)}] \in \mathbb{R}^{|d(S+1)|} \tag{6.13}$$

In these formulations, $v_i^{(s)}$ denotes the embedding of the i-th agent at scale s, and $|d(S+1)|$ represents the dimensionality of the concatenated embeddings, accounting for all $S+1$ scales.

6.2 Posterior Distribution Learner

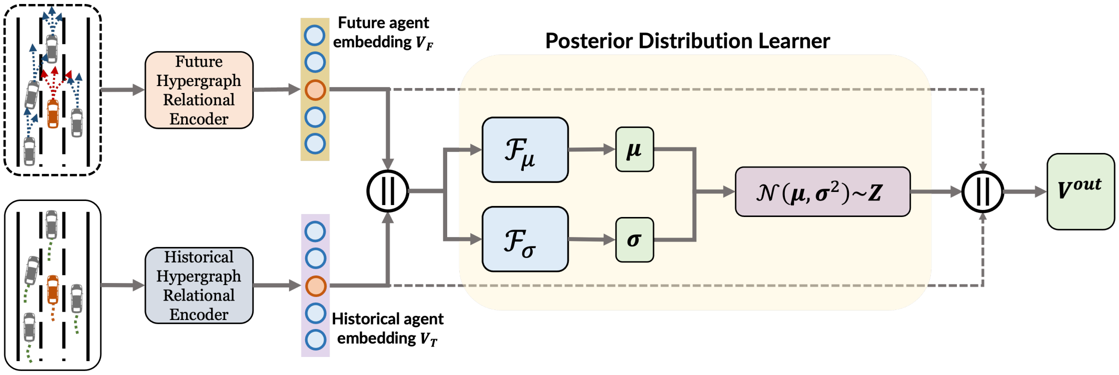

Future Hypergraph Relational Encoder
Future agent embedding V_F
Posterior Distribution Learner
$\mathcal{F}_\mu$
μ
$\mathcal{F}_\sigma$
σ
$\mathcal{N}(\mu, \sigma^2) \sim Z$
V^{out}
Historical Hypergraph Relational Encoder
Historical agent embedding V_T

Figure 6.4: Posterior Distribution Learner.

In our study, we incorporated multi-scale hypergraph embeddings into a multi-agent trajectory generation system using the CVAE framework [112] to address the stochastic nature of each agent's behavior. Let $\log p(X_F|X_T)$ denote the log-likelihood of predicted future trajectories X_F given past trajectories X_T. The corresponding evidence lower bound (ELBO) is defined as follows:

$$\begin{aligned} \log p(X_F|X_T) \geq \mathbb{E}_{q(\mathbf{Z}|X_F,X_T)} &\log p(X_F|\mathbf{Z}, X_T) \\ &- \mathrm{KL}(q(\mathbf{Z}|X_F, X_T) \parallel p(\mathbf{Z}|X_T)), \end{aligned} \tag{6.14}$$

where $\mathbf{Z} \in \mathbb{R}^{N \times d_z}$ represents the latent codes corresponding to all agents; $p(\mathbf{Z}|X_T)$ is the conditional prior of $\mathbf{Z}$, modeled as a Gaussian distribution. In this framework, $q(\mathbf{Z}|X_F, X_T)$ is implemented through an encoding process for embedding learning, and $p(X_F|\mathbf{Z}, X_T)$ is realized via a decoding process that forecasts the future trajectories X_F.

Thus, the goal of the Posterior Distribution Learner is to derive the Gaussian parameters for the approximate posterior distribution. This entails calculating the mean μ_q and the variance σ_q from the final output embeddings V_F and the target embeddings V_T. These parameters are produced using two distinct trainable multilayer perceptrons (MLPs), denoted as $\mathcal{F}_\mu$ and $\mathcal{F}_\sigma$. The latent code Z, which represents potential trajectories, is then sampled from a Gaussian distribution characterized by these means and variances. The final output embeddings V_{out} are formed by concatenating the latent code Z with the final output embeddings V_F and the target embeddings V_T. The equations that describe these processes are as follows:

$$\mu_q = \mathcal{F}_\mu(V_F, V_T) \tag{6.15}$$

$$\sigma_q = \mathcal{F}_\sigma(V_F, V_T) \tag{6.16}$$

$$Z \sim \mathcal{N}(\mu_q, \mathrm{Diag}(\sigma_q^2)) \tag{6.17}$$

$$V_{out} = [Z, V_F, V_T] \tag{6.18}$$

In these notations, μ_q and σ_q represent the mean and variance of the approximated posterior distribution. $\mathcal{F}_\mu$ and $\mathcal{F}_\sigma$ are the trainable MLPs that produce these parameters. Z denotes the latent code of possible trajectories, and $Vout$ stands for the output embeddings, which is an informative fusion encapsulating the potential future states as predicted by the model.

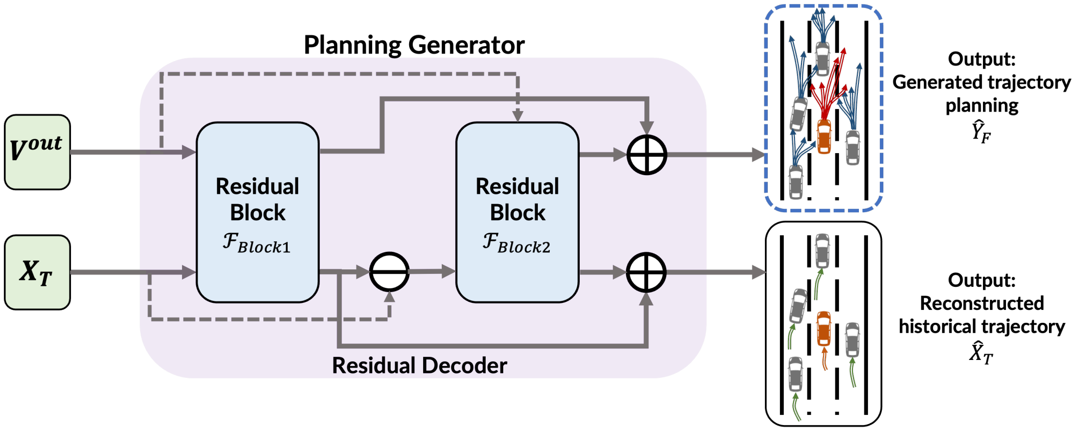

Planning Generator
V^{out}
X_T
Residual Block $\mathcal{F}_{Block1}$
Residual Block $\mathcal{F}_{Block2}$
Residual Decoder
Output: Generated trajectory planning $\hat{Y}_F$
Output: Reconstructed historical trajectory $\hat{X}_T$

Figure 6.5: Residual Decoder.

6.3 Residual Decoder

The Residual Decoder is designed with a dual objective: predicting future trajectories and reconstructing past trajectories from the given embeddings. This decoder employs successive processing blocks, each contributing a residual that refines the trajectory estimates. The first processing block, $\mathcal{F}_{Block1}$, takes the output embeddings V^{out} and the target past trajectory X_T to generate initial estimates of the future and reconstructed past trajectories, denoted as $\hat{X}_{F,1}$ and $\hat{X}_{T,1}$ respectively.

$$\hat{X}_{F,1}, \hat{X}_{T,1} = \mathcal{F}_{Block1}(V^{out}, X_T) \tag{6.19}$$

Next, the second block, $\mathcal{F}_{Block2}$, refines these estimates by considering the output embeddings and the residual of the past trajectory, defined as the difference between the target past trajectory and the initial reconstructed past trajectory $X_T - \hat{X}_{T,1}$. This process yields the second set of refined residuals, $\hat{X}_{F,2}$ and $\hat{X}_{T,2}$:

$$\hat{X}_{F,2}, \hat{X}_{T,2} = \mathcal{F}_{Block2}(V^{out}, X_T - \hat{X}_{T,1}) \tag{6.20}$$

Both $\mathcal{F}_{Block1}$ and $\mathcal{F}_{Block2}$ are composed of a GRU encoder for sequence encoding and two MLPs serving as the output header. The final predicted future trajectory $\hat{X}_F$ and the reconstructed past trajectory $\hat{X}_T$ are obtained by summing the respective residuals from both processing blocks:

$$\hat{X}_F = \hat{X}_{F,1} + \hat{X}_{F,2} \tag{6.21}$$

$$\hat{X}_T = \hat{X}_{T,1} + \hat{X}_{T,2} \tag{6.22}$$

This iterative refinement approach leverages the deep learning model's capacity to capture complex patterns in the data, enhancing the accuracy of both predictions and reconstructions. The Residual Decoder ultimately generates a set of K planned trajectories for each vehicle in the vehicle group.

The Residual Decoder generates number of K planned trajectories for each vehicle in the vehicle group.

6.4 Neural Planner and Refinement

To ensure robust motion planning for the target vehicle, based on the multi-modal initial planning results and the generated motion states of surrounding vehicles, we adopt a prediction-guided pipeline for planning refinement. Initially, the best-generated trajectory is selected, followed by planning optimization to refine this trajectory.

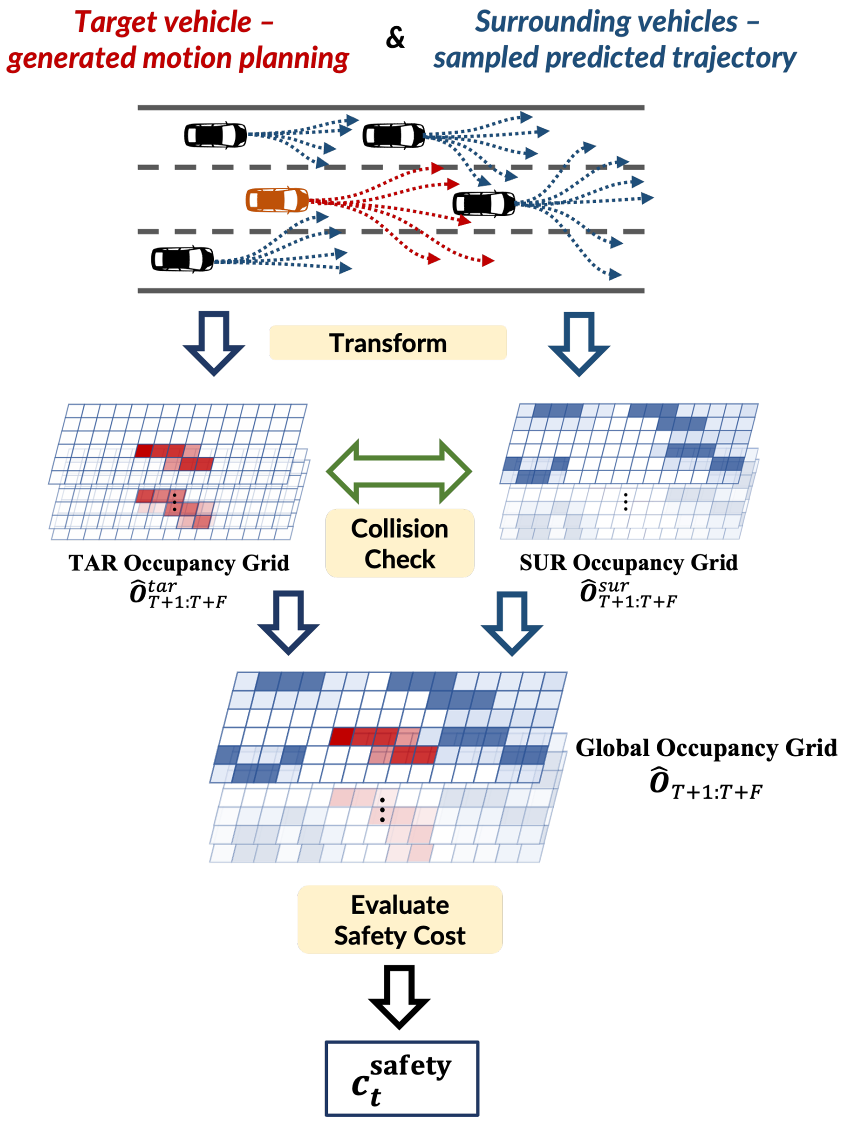


Figure 6.6: Prediction guided safety evaluation workflow.

To enhance safety and motion performance, we evaluate and select the most optimal trajectory for navigation. The objective function for this process is defined as:

$$\tau^* = \underset{\tau \in \Lambda}{\arg\min} C(\tau, \hat{Y}^{(K,N)}_{T+1:T+F}) \tag{6.23}$$

where the cost function C for a given trajectory τ and predicted states $\hat{Y}^{(K,N)}_{T+1+F}$ is the sum of individual cost components at each step t, weighted by their respective importance:

$$C(\tau, \hat{Y}^{(K,N)}_{T+1:T+F}) = \sum_i \omega_i \sum_t c^i_t \tag{6.24}$$

Here, c^i_t represents the cost of type i at step t, encompassing efficiency, comfort, and safety over the prediction horizon from $T+1$ to $T+F$, with ω_i as the weight for each cost type i.

A candidate planned trajectory τ comprises positions (x_t, y_t) over the time steps of interest:

$$\tau = \{(x_t, y_t) | \forall t \in [T+1, \ldots, T+F]\} \tag{6.25}$$

The efficiency cost aims to maintain a consistent longitudinal speed within speed limits, thereby enhancing on-road progress. This is crucial for ensuring the vehicle reaches its destination in a timely manner while adhering to traffic regulations. The efficiency cost is mathematically expressed as:

$$c^{efficiency}_t = \dot{x}_t - v_{limit} \tag{6.26}$$

The comfort cost penalizes fluctuations in longitudinal acceleration, jerk, and lateral acceleration. This is important for providing a smooth driving experience, reducing the likelihood of abrupt movements that could discomfort passengers or cause wear on vehicle components. The comfort cost is defined as:

$$c_t^{comfort} = \ddot{x}_t + \ddot{y}_t + \dddot{x}_t \tag{6.27}$$

Safety is addressed by performing a collision check between the predicted occupancy grids for the target and surrounding vehicles. If a collision is detected, the trajectory in question is discarded. The Potential Risks safety evaluation metric, introduced by [109], computes the regulated resultant force between the target vehicle and other vehicles as a measure of potential risks. This ensures that the selected trajectory minimizes the risk of accidents and maintains a safe distance from other vehicles. The safety cost is formulated as:

$$PR_t^{(tar,j)} = \tanh\left(\sqrt{\left(\frac{F_{t,x}^{(tar,j)}}{\sigma_{Fx}}\right)^2 + \left(\frac{F_{t,y}^{(tar,j)}}{\sigma_{Fy}}\right)^2}\right) \tag{6.28}$$

The safety cost aggregates the potential risks associated with all other vehicles:

$$c_t^{safety} = \sum_{j \in [1,N]} PR_t^{(tar,j)} \tag{6.29}$$

This comprehensive cost evaluation ensures that the vehicle's planned trajectory optimizes for efficiency and comfort while maintaining the highest safety standards.

Chapter 7

Experiment Settings

7.1 Data Preparations

This research utilizes two well-known datasets for training and validating the model: the Next Generation Simulation (NGSIM) dataset [113, 114] and the HighD dataset [115].

The NGSIM dataset provides an extensive collection of vehicle trajectory data, capturing traffic activity on the eastbound I-80 in the San Francisco Bay area and the southbound US 101 in Los Angeles. Compiled by the U.S. Department of Transportation in 2015, this dataset includes real-world highway scenarios recorded using overhead cameras at a sampling rate of 10 Hz.

In contrast, the HighD dataset is based on aerial drone recordings conducted at a frequency of 25 Hz between 2017 and 2018 near Cologne, Germany. Covering approximately 420 meters of bidirectional roadways, it documents the movements of around 110,000 vehicles, including cars and trucks, traveling a cumulative distance of 45,000 km.

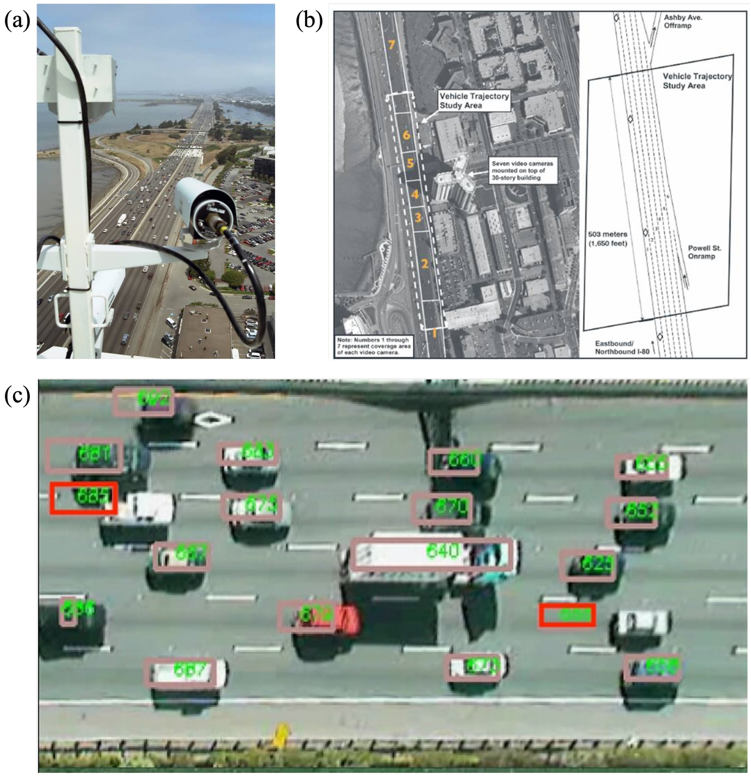
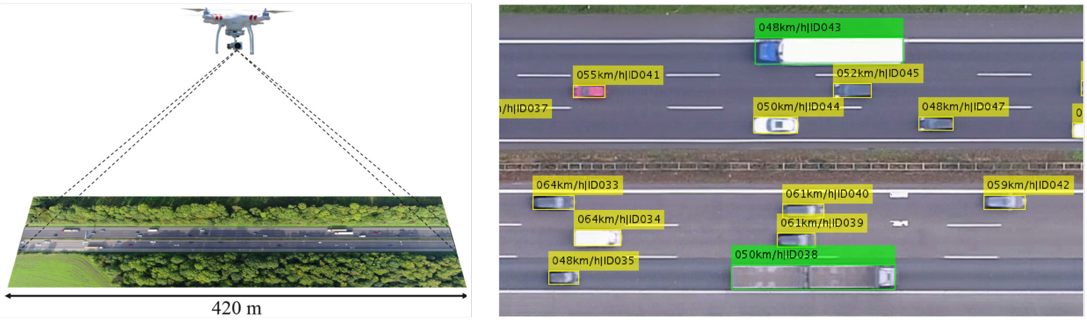

(a)
(b)
Ashby Ave. Offramp
Vehicle Trajectory Study Area
Vehicle Trajectory Study Area
Seven video cameras mounted on top of 30-story building
503 meters (1,650 feet)
Powell St. Onramp
Eastbound/ Northbound I-80
Note: Numbers 1 through 7 represent coverage area of each video camera.
7
6
5
4
3
2
1
(c)
640

Figure 7.1: NGSIM Dataset. (a) A digital video camera recording vehicle trajectory data I-80. (b) I-80 study area. (c) Recorded vehicles with bounding boxes. and HighD dataset (right).

After pre-processing, the NGSIM dataset includes 662,000 rows of data, capturing 1,380 individual trajectories. The HighD dataset contains 1.09 million data entries, covering 3,913 individual trajectories. For model training and evaluation, the data is divided into 70% for the training set and 30% for the test set. The temporal parameters for the model are set to $T = 30$ frames for the historical horizon and $F = 50$ frames for the prediction horizon.



(a) (b)

Figure 7.2: HighD Dataset. (a) A drone capturing traffic from a bird's eye view on a road section with a length of about 420 m. (b) Recorded vehicles with bounding boxes.

7.2 Training Metrics

Training Metrics of GTIMP Model and GIRAFFE Model. This research adopts a two-stage training strategy, leveraging various loss functions to optimize the learning efficiency of neural network parameters.

In the first phase, spanning five epochs, the goal is to minimize the Mean Square Error (MSE) loss, a common metric for trajectory prediction problems:

$$MSE(\hat{Y};Y) = \frac{1}{F}\sum_{t'=T+1}^{T+F}\left((\mu_{t',x} - x_{t'})^2 + (\mu_{t',y} - y_{t'})^2\right) \tag{7.1}$$

where $Y = \{(x_{t'}, y_{t'})\}$ represents the ground truth position of the target vehicle at timestamp t', $\hat{Y} = \{(\mu_{t',x}, \mu_{t',y})\}$ denotes the predicted position.

From the sixth epoch onwards, the focus transitions to enhancing the network's capabilities via the negative log-likelihood (NLL) loss:

$$NLL(\hat{Y};Y) = -\log\left(\sum_m P_\theta(Y|X,m)P(m|X)\right) \tag{7.2}$$

For each training instance, a single intention category is designated, steering the optimization process towards this variant of the NLL loss function:

$$\begin{aligned} NLL(\hat{Y};Y) &= -\log(P_{\theta}(Y|X,m)P(m|X)) \\ &= -\log P_{\theta}(Y|X,m) - \log P(m|X) \\ &= NLL_{traj}(\hat{Y};Y) + NLL_{int}(\hat{M};M) \end{aligned} \tag{7.3}$$

The prediction of intentions is treated as a classification problem along the future time horizon, employing the cross-entropy loss for NLL_{int}:

$$NLL_{int}(\hat{M};M) = -\sum_{m\in M} m \log P(\hat{m}|X) \tag{7.4}$$

where m denotes the ground truth intention. For the trajectory-related term NLL_{traj}, based on the bivariate Gaussian distribution, we have:

$$\begin{aligned} & NLL_{traj}(\hat{Y};Y) \\ & = \sum_{t'=T+1}^{T+F} \Bigg(\log \Bigg(2\pi\sigma_{t',x}\sigma_{t',y}\sqrt{1-\rho_{t'}^{2}} \\ & \quad + \frac{1}{2(1-\rho_{t'}^{2})} \Bigg(\frac{(\mu_{t',x}-x_{t'})^{2}}{\sigma_{t',x}^{2}} \\ & \quad - \frac{(\mu_{t',x}-x_{t'})(\mu_{t',y}-y_{t'})}{\sigma_{t',x}\sigma_{t',y}} + \frac{(\mu_{t',y}-y_{t'})^{2}}{\sigma_{t',y}^{2}} \Bigg) \Bigg) \Bigg) \end{aligned} \tag{7.5}$$

Consequently, the overall two-stage loss function $\mathcal{L}$ is formulated as:

$$\mathcal{L} = \begin{cases} MSE(\hat{Y};Y) + \alpha \cdot NLL_{int}(\hat{M};M) & \\ \quad +\beta \cdot MSE(\tilde{H}_F;H_F), & \text{if epoch} \leq 5 \\ NLL_{traj}(\hat{Y};Y) + \alpha \cdot NLL_{int}(\hat{M};M) & \\ \quad +\beta \cdot MSE(\tilde{H}_F;H_F), & \text{o.w.} \end{cases} \tag{7.6}$$

where Y and $\hat{Y}$ represent the ground truth and predicted position distributions of the target vehicle, respectively. Similarly, M and $\hat{M}$ denote the ground truth and predicted intentions. Additionally, H_F and $\tilde{H}_F$ represent the ground truth and predicted values corresponding to the future positions of all vehicles in the vehicle group, including the target and surrounding vehicles.

The proposed network architecture is implemented using the PyTorch deep learning framework. The Adam optimizer is employed with an initial learning rate of 0.001 and a decay factor to train the network in an end-to-end manner. The parameter settings are presented in Table 7.2.

Table 7.1: Hyperparameter Settings of GIMTP and GIRAFFE

Parameter	Value	Parameter	Value
T	30	neuron # of GRU	128
F	50	neuron # of MLP$_1^{MD}$	128
neuron # of DGCN	256	neuron # of MLP$_2^{MD}$	128
neuron # of MLP$_1^{IP}$	256	learning rate	0.01
neuron # of MLP$_2^{IP}$	256	decaying factor	0.95
neuron # of LatMLP	256	α	0.2
neuron # of LonMLP	256	β	0.1

Training Metrics of RHINO Model. The training loss of RHINO consists of three components:

The Evidence Lower Bound (ELBO) Loss: This is a standard loss used in variational autoencoders [112]. It consists of a reconstruction loss, which measures the discrepancy between the predicted and actual future trajectories, and a regularization term, expressed as the Kullback-Leibler (KL) divergence between the learned posterior distribution and a prior distribution. This regularization helps the learned distribution approximate the prior distribution, improving generalization.

$$\mathcal{L}_{elbo} = \alpha \|\hat{X}_F - X_F\|_2^2 + \beta \mathrm{KL}\left(\mathcal{N}\left(\mu_q, \mathrm{Diag}(\sigma_q^2)\right) \middle\| \mathcal{N}\left(0, \lambda I\right)\right) \tag{7.7}$$

The Historical Trajectory Reconstruction Loss: This loss measures the accuracy of the model in reconstructing the historical trajectories of vehicles. By minimizing this loss, the model is trained to accurately recall past states, which is crucial for understanding the dynamics and behaviors that lead to future states.

$$\mathcal{L}_{recon} = \gamma \|\hat{X}_T - X_T\|_2^2 \tag{7.8}$$

The Variety Loss: Inspired by Social-GAN [71], this component ensures diversity in the predicted future trajectories. It does so by minimizing the error across multiple sampled trajectories, thereby encouraging the model to generate a wide range of plausible future scenarios, which is vital in the context of multi-agent systems where uncertainty and variability are inherent.

$$\mathcal{L}_{variety} = \min_k \|\hat{X}_F^{(k)} - X_T\|_2^2 \tag{7.9}$$

The total training loss function is a weighted sum of these components:

$$\mathcal{L} = \alpha \mathcal{L}_{elbo} + \beta \mathcal{L}_{recon} + \gamma \mathcal{L}_{variety} \tag{7.10}$$

Each term in the loss function serves a specific purpose. The coefficient α scales the ELBO loss to ensure the model balances between accurate future trajectory prediction and adherence to the prior distribution. The coefficient β scales the reconstruction loss to enforce the model's ability to recall and utilize historical data effectively. The coefficient γ scales the variety loss to promote the generation of diverse and realistic future trajectories. The parameter λ in the KL divergence term controls the prior distribution's influence on the learned posterior distribution.

Table 7.2: Hyperparameter Settings of RHINO

Parameter	Value	Parameter	Value
T	30	decaying factor	0.6
nF	50	α	1
neuron # of MLPs	128	β	0.8
learning rate	0.001	γ	0.5

7.3 Evaluation Metrics

To measure the predictive accuracy of the model, we use the Root Mean Square Error (RMSE). This metric quantifies the deviation between the predicted position, represented as $(\mu^l_{t',x}, \mu^l_{t',y})$, and the ground truth position, signified by $(x^l_{t'}, y^l_{t'})$. Such evaluations are undertaken across discrete temporal markers t' encompassed within the predictive horizon prediction horizon $[T+1, T+F]$.

$$RMSE = \sqrt{\frac{1}{LF}\sum_{l=1}^{L}\sum_{t'=T+1}^{T+F}\Big((\mu^l_{t',x} - x^l_{t'})^2 + (\mu^l_{t',y} - y^l_{t'})^2\Big)} \tag{7.11}$$

where the superscript l is employed to signify the l-th test sample from the aggregate test sample set with length L. In contexts entailing multi-modal predictions, our proposed model is capable of rendering multiple trajectory outputs. Amongst these, the trajectory bearing the preeminent probability is harnessed for RMSE computation. Conversely, in more conventional settings, our model yields a single trajectory, which is then appropriated for the ensuing evaluation.

7.4 Baseline Models

We compare our proposed model with the following baseline models:

- Social-LSTM (S-LSTM) [17]: his model uses a shared LSTM to encode the raw trajectory data for each vehicle and aggregates the features of different vehicles through a social pooling layer.

- Convolutional Social-LSTM (CS-LSTM) [42]: Unlike S-LSTM, this model captures social interactions by stacking convolutional and pooling layers and accounts for multi-modality based on the predicted intention.

- Planning-informed prediction (PiP) [87]: This model integrates trajectory prediction with the planning of the target vehicle by conditioning on multiple candidate trajectories.

- Graph-based Interaction-aware Trajectory Prediction (GRIP) [18]: This model employs a graph-based representation for interactions between objects, uses graph convolutional layers for feature extraction, and implements an encoder-decoder LSTM for predictive analysis.

Chapter 8

Experiment Results and Discussion

In this chapter, we present the results of our experiments conducted using two prominent datasets for trajectory prediction. We evaluate the performance of four baseline models in comparison to our proposed framework, employing specific evaluation metrics within a three-lane highway scenario. Additionally, we perform ablation studies to highlight the importance of each model component and to provide deeper insights into the design of our model.

8.1 Model Performance Comparison

Performance Comparison of GIMTP Model. The comparative results are presented in Table 8.1 and Figure 8.1. Our proposed framework exhibits superior performance, as measured by Root Mean Square Error (RMSE), across a prediction horizon of 50 frames when compared to the existing baseline models. Specifically, it achieves a lower prediction error than S-LSTM, CS-LSTM, PiP, and GRIP. These results underscore the effectiveness of our model in capturing essential features necessary for accurate long-term trajectory predictions.

Table 8.1: Prediction Error Obtained by Different Models in RMSE

Dataset	Horizon (Frame)	S-LSTM	CS-LSTM	PiP	GRIP	**GIMTP**
NGSIM	10	0.65	0.61	0.55	0.37	**0.35**
	20	1.31	1.27	1.18	0.86	**0.82**
	30	2.16	2.08	1.94	1.45	**1.39**
	40	3.25	3.10	2.88	**2.21**	2.24
	50	4.55	4.37	4.04	3.16	**3.05**
HighD	10	0.22	0.22	0.17	0.29	**0.17**
	20	0.62	0.61	0.52	0.68	**0.39**
	30	1.27	1.24	1.05	1.17	**0.73**
	40	2.15	2.10	1.76	1.88	**1.02**
	50	3.41	3.27	2.63	2.76	**1.42**

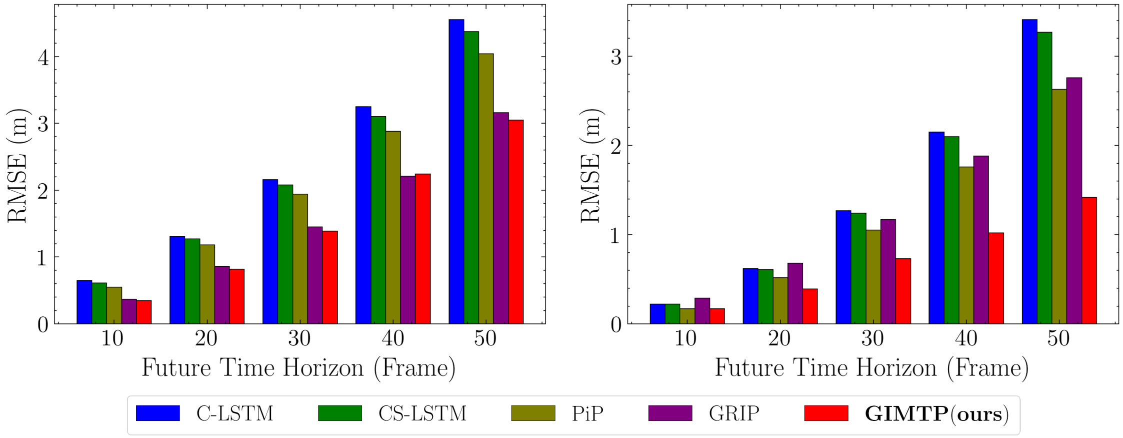


Figure 8.1: Prediction error obtained by different models in RMSE on NGSIM dataset (left) and HighD dataset (right).

Overall, our proposed framework not only surpasses the baseline models on the HighD dataset but also demonstrates commendable performance on the NGSIM dataset. These findings validate the robustness and reliability of our approach in diverse traffic scenarios.

Performance Comparison of RHINO Model. Since the RHINO framework incorporates the enhanced GIRAFFE, we compare the trajectory generation capabilities of RHINO-Gen with the original GIMTP and the enhanced GIRAFFE. Both the RHINO-Gen model and the enhanced GIRAFFE model consistently outperform the baseline models, demonstrating superior performance across various metrics. This suggests that our proposed approaches effectively address the limitations of traditional models, providing a robust framework for capturing complex interactions and generating accurate predictions.

Table 8.2: Generation Error Obtained by Different Models in RMSE

Dataset	Horizon (Frame)	GIMTP	**GIRAFFE**	**RHINO**
NGSIM	10	0.35	0.38	**0.32**
	20	0.82	0.89	**0.78**
	30	1.39	1.45	**1.34**
	40	2.24	2.46	**2.17**
	50	3.05	3.24	**2.97**
HighD	10	0.17	0.19	**0.19**
	20	0.39	0.42	**0.26**
	30	0.73	0.81	**0.42**
	40	1.02	1.13	**0.65**
	50	1.42	1.56	**0.89**

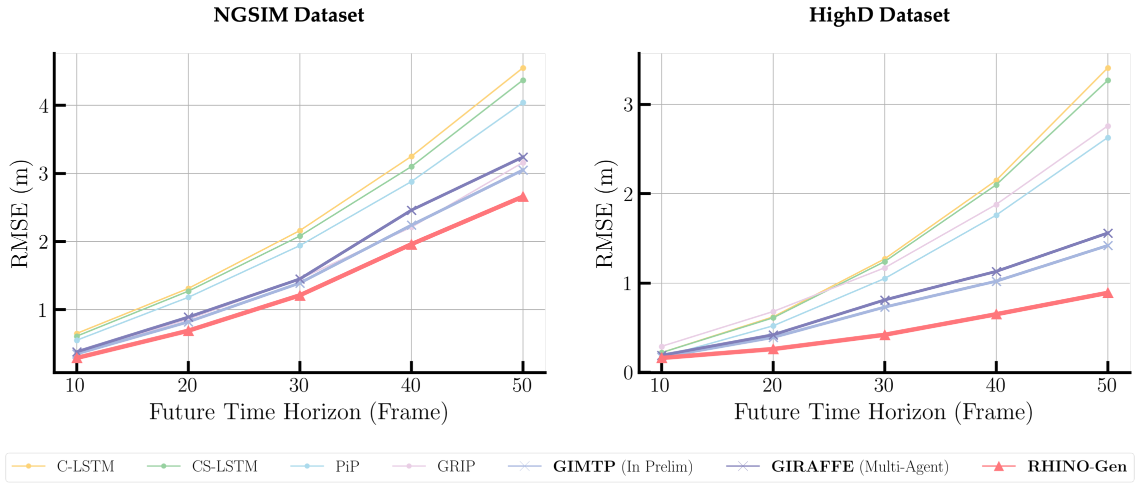


Figure 8.2: Generation error obtained by different models in RMSE on NGSIM dataset (left) and HighD dataset (right).

8.2 Results of Multi-modal Predictions

We begin by evaluating the performance of the multi-modal prediction approach. The results of experiments involving the prediction of multiple trajectories, considering different lateral intentions using the HighD dataset, are displayed in Figure 8.3. The green line indicates the ground truth trajectory, while the solid red line represents the predicted trajectory using the fused intention features. Potential trajectories, influenced by the application of a lateral intention from the current time step, are depicted by the purple lines.

In this experiment, the probabilities of Lane Keeping (LK), Left Lane Change (LLC), and Right Lane Change (RLC) intentions are fixed at 1 throughout the entire future time horizon, enforcing mandatory lane changing or lane keeping for the predicted trajectories. The color opacity signifies the probability of lane-changing intentions at the current time step T, with more solid colors indicating higher probabilities, and lighter colors representing lower probabilities.

Figures 8.3 illustrate the longitudinal and lateral positions over time, respectively. Figure 8.4 showcases the prediction results of the target vehicles along with the historical trajectories of the surrounding vehicles across six experiments in three-lane highway scenarios, with the dashed grey lines indicating the lane markings' positions.

Our model demonstrates proficiency in predicting the probabilities associated with each lateral and longitudinal intention while concurrently forecasting the corresponding future trajectories for each intention. Additionally, the model allows for manual inputs and adjustments of the intentions and their respective probabilities at each time step within the future horizon F, enabling the observation of the corresponding generated trajectories. This flexibility enhances the evaluation of the model's performance across various scenarios.

8.3 Results of Trajectory, Intention, and Interaction Prediction

We conducted a series of experiments to analyze the predicted trajectories, intentions, and interactions of vehicles, as illustrated in Figure 8.5. Figures 8.5(b) and 8.5(e) specifically depict the predicted intentions of the target vehicle at each time step within the future time horizon for different lane-changing scenarios. Figure 8.5(b) demonstrates the process of a right lane change, while Figure 8.5(e) focuses

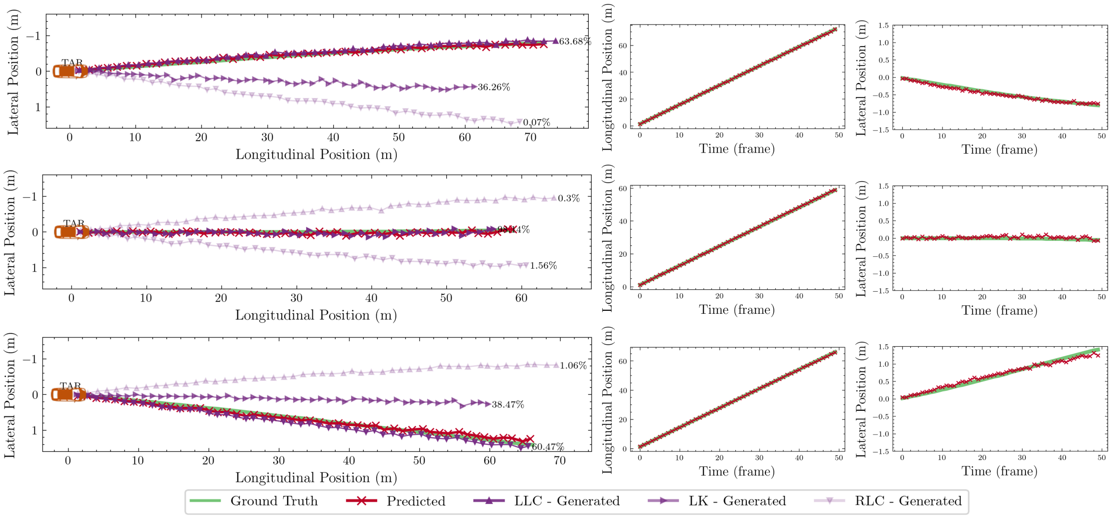


Figure 8.3: Results of multi-modal trajectory generation.

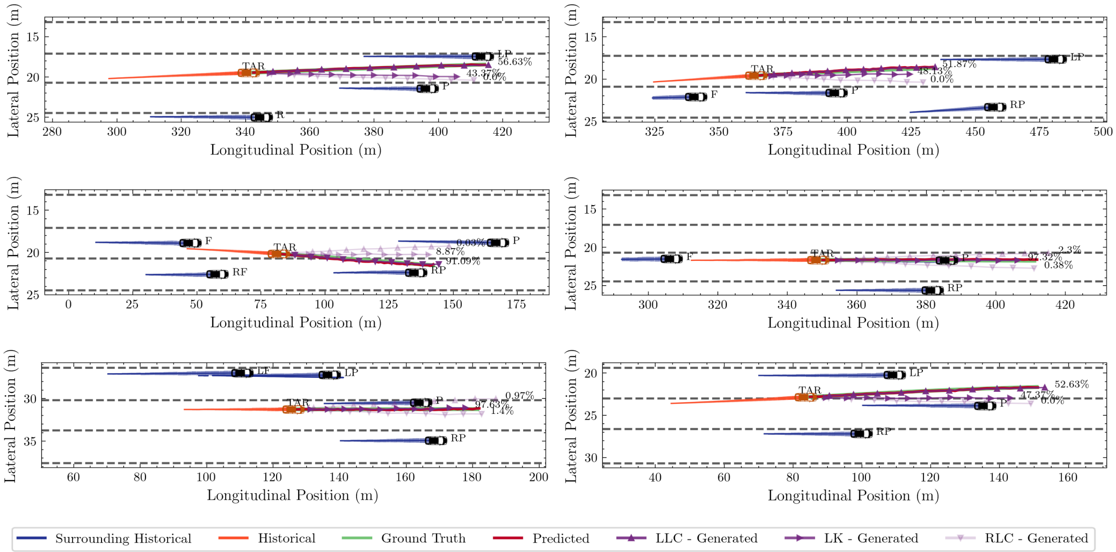


Figure 8.4: Results of multi-modal trajectory prediction experiments in three-lane highway scenarios.

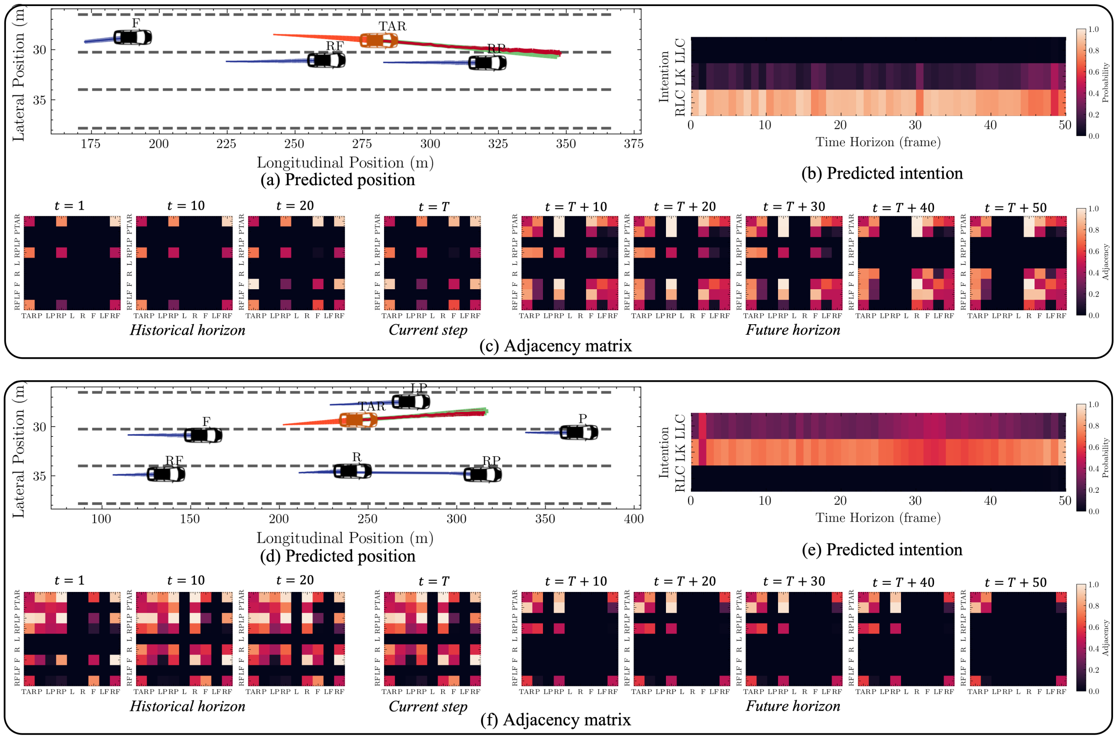
on lane-keeping scenarios.

Additionally, Figures 8.5(c) and 8.5(f) present the adjacency matrices of the vehicle group, which are generated by the historical graph embedding $\tilde{H}_T$ and the future-guided graph embedding $\tilde{H}_F$ within the DGCN encoder of the Graph Interaction Encoder. These adjacency matrices reflect vehicle interactions at the current time step, with lighter colors indicating more significant interactions.

The future-guided adjacency matrix generated by the model effectively predicts vehicles' entry into and exit from the vehicle group, along with their respective motion states. This future-guided adjacency matrix is used as part of the input for the Intention Predictor to create fused hidden states for the future horizon. The proposed model demonstrates a successful integration of accurate future trajectory predictions with future intention estimations, highlighting its capability to foresee vehicle interactions and motion states accurately.

8.4 Results of Planning Trajectory Generation

The experimental results for planning trajectory generation of the top K trajectories using the HighD dataset are presented in Figure 8.6. The historical trajectory is depicted by the orange dashed line, while the green line represents the ground truth future trajectory. The potential trajectories generated by `RHINO-Gen` are shown in blue, with the best generation highlighted by the solid red line. `RHINO-Gen` demonstrates strong generative capabilities, effectively producing plausible motion planning scenarios. This proficiency is crucial for applications that require anticipat-



(a) Predicted position

(b) Predicted intention

(c) Adjacency matrix

(d) Predicted position

(e) Predicted intention

(f) Adjacency matrix

Figure 8.5: Results of trajectory prediction experiment. (a) and (d) Predicted trajectory v.s. ground truth trajectory. (b) and (e) Predicted lateral intention with probability. (c) and (f) Adjacency matrix of the vehicle group at historical and future time steps.

ing multiple potential future states. The generative approach not only enhances predictive accuracy but also provides valuable insights into possible future trajectories, facilitating more informed decision-making. The diversity of generated trajectories showcases the model's ability to account for different potential driving behaviors, thereby improving the robustness of motion planning.

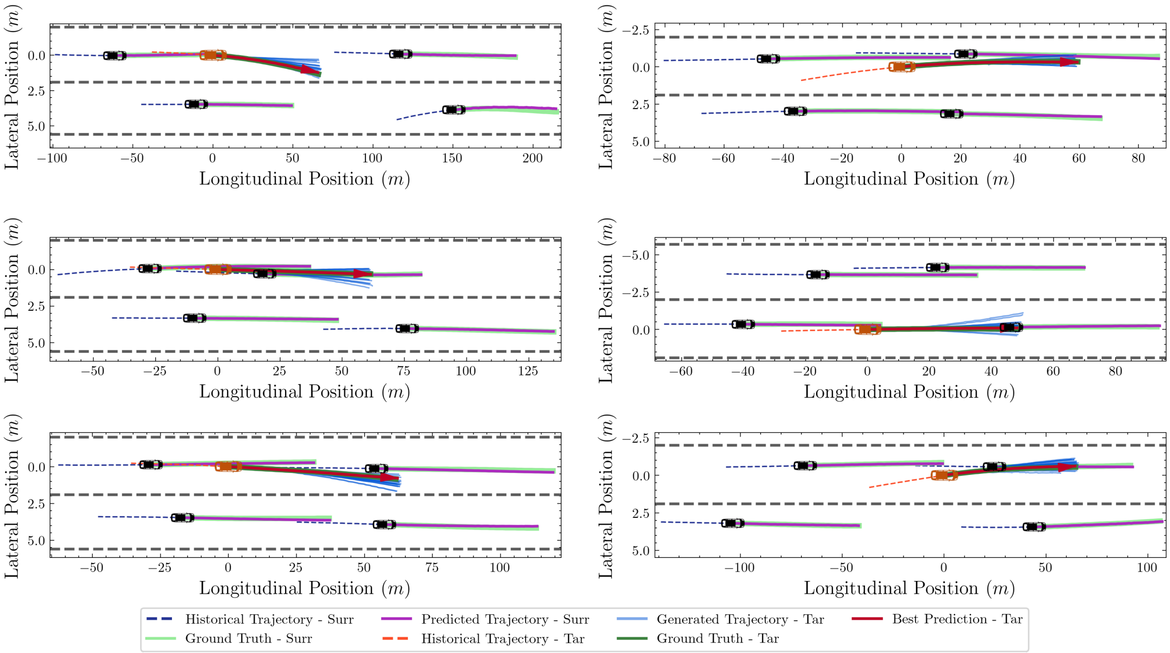


Figure 8.6: Results of trajectory planning generation in three-lane highway scenarios.

8.5 Results of Planning Trajectory Error Analysis

The analysis of trajectory generation inaccuracies is illustrated in Figure 8.7. The generated trajectories in both the longitudinal and lateral axes, along with the error box plots, are displayed. The box plot reveals that errors in both axes increase with the prediction time step. However, the errors remain within an acceptable range, indicating decent model performance. These findings demonstrate high precision in trajectory generation when compared with the ground truth future trajectory. Notably, the model maintains a lower error margin for shorter prediction horizons, which is critical for short-term planning and reactive maneuvers in dynamic traffic scenarios. The consistency of the error growth pattern suggests that the model's predictive capability degrades gracefully over longer horizons, maintaining practical utility in real-world applications.

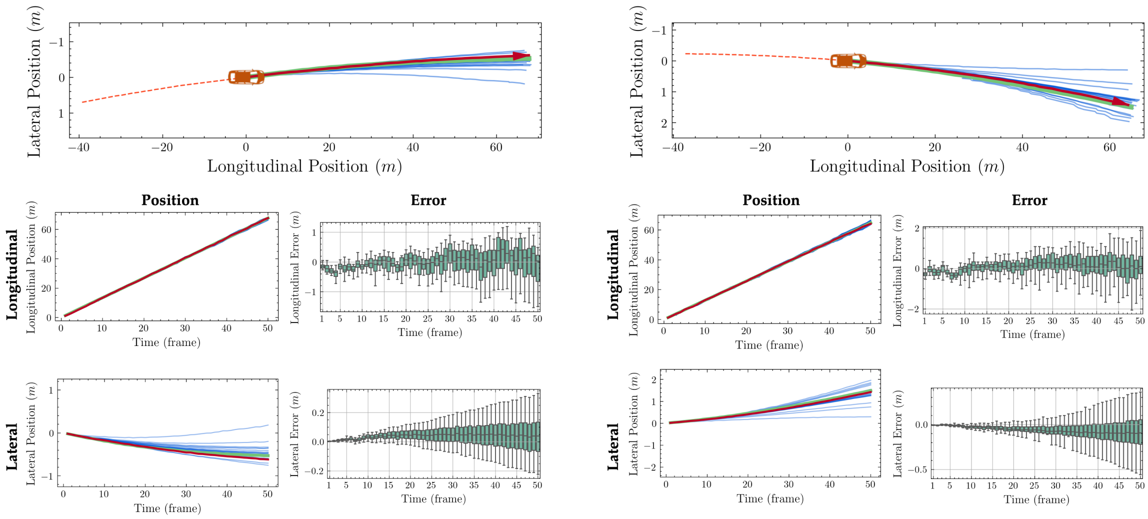


Figure 8.7: Results of longitudinal and lateral error in trajectory generation planning.

8.6 Results of Planning Trajectory Refinement

`RHINO-Plan` enhances the framework by refining trajectory planning for the target vehicle, with a focus on optimizing safety, efficiency, and comfort. This prediction-guided refinement ensures that the generated plans are not only feasible but also optimized for practical application, addressing critical concerns in real-world deployment.

Figure 8.8 illustrates the performance comparison of planning trajectories. Multiple plausible trajectories are generated for all vehicles in the scenario, with the

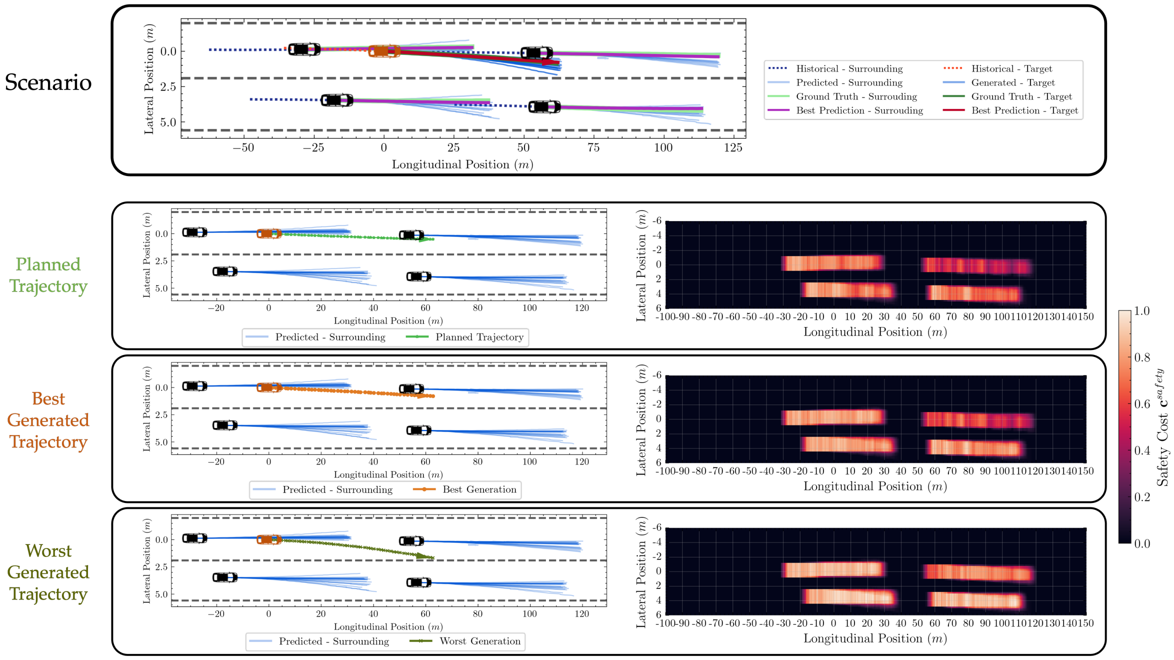


Figure 8.8: Results of planning trajectory performance comparison.

refined planned trajectory contrasted against the best and worst generated trajectories. The heatmaps in the right columns display the safety cost c^{safety} for the corresponding trajectory planning, indicating a lower safety cost for the refined planned trajectory and a higher cost for the worst generated trajectory. This comparison underscores the effectiveness of `RHINO-Plan` in optimizing trajectories for safety, thereby reducing potential risks in complex traffic environments.

Figure 8.9 presents the results of planning trajectory optimization and refinement. The vehicle dynamics parameters, including heading angle, longitudinal and lateral positions, velocity, and acceleration of the refined motion planning and candidate planning trajectories, are displayed. The heatmap shows the cost values of the 20 candidate trajectory planning. Initially, the motion planning for the target

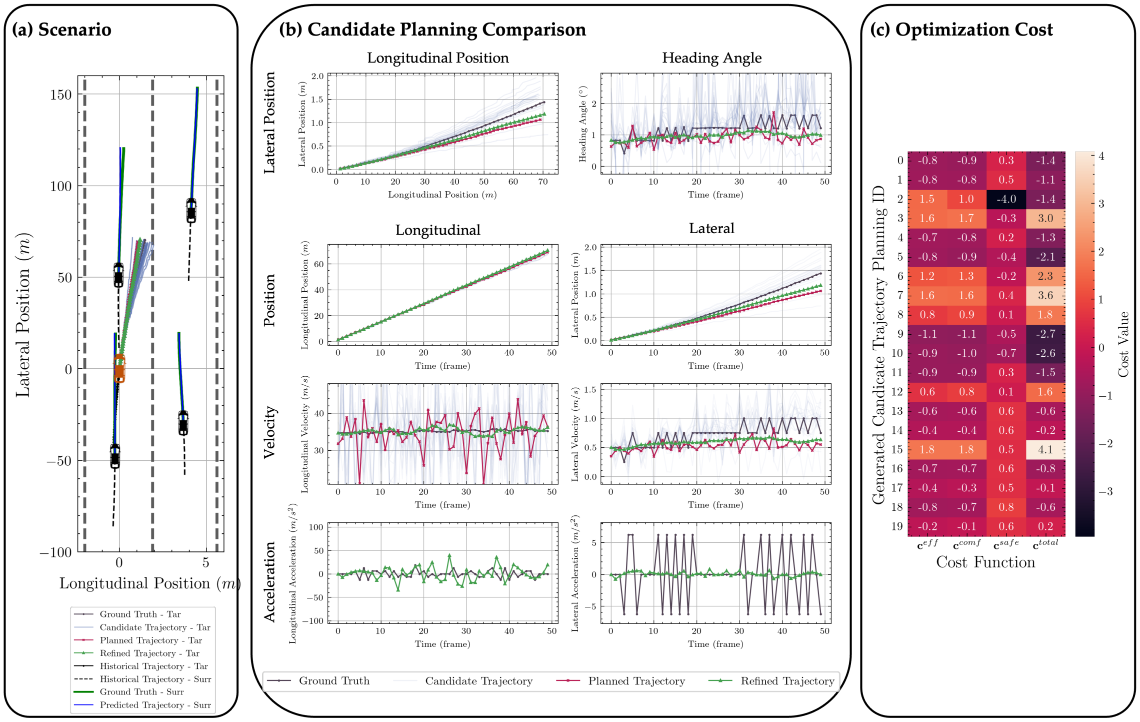


Figure 8.9: Results of planning trajectory optimization and refinement.

vehicle generates positions that may amplify inaccuracies during derivative calculations, rendering them infeasible for vehicle controllers. The best motion generation, characterized by the minimum total cost—encompassing safety, efficiency, and comfort—is selected. The planning refinement further smooths the best trajectory, ensuring that the vehicle dynamics, particularly acceleration and velocity, are brought to a feasible level.

This refinement process is critical for ensuring that the planned trajectories are not only theoretically optimal but also practically executable by real-world vehicle controllers. The optimization process includes fine-tuning the trajectory to adhere to realistic vehicle dynamics constraints, such as maximum acceleration and deceleration limits, ensuring smooth transitions and maintaining passenger comfort. By prioritizing these aspects, RHINO-Plan enhances the overall reliability and performance of the system, leading to human-like motion planning that is both safe and efficient.

8.7 Ablation Study

Abalation Study of GIMTP Model. An ablation study was conducted to gain deeper insights into the performance of the GIMTP model, particularly the impact of its various components on prediction accuracy. This was achieved by selectively disabling specific components from the GIMTP architecture. The study considered the following four variants:

- **GIMTP w/o PR**: This variant excludes the adjacency matrix influenced by potential risks, relying solely on neighborhood and distance metrics within the Dynamic Graph Embedding Module.

- **GIMTP w/o DGCN**: This variant replaces the Diffusion Graph Convolutional Network (DGCN) architecture in the Interaction Encoder with a simpler Graph Convolutional Network (GCN).

- **GIMTP w/o FG**: This variant omits the prediction of the encoded future-guided graph matrix that represents future states within the Interaction Encoder, using only historical motion states for intention prediction and feature fusion.

- `GIMTP w/o FF`: This variant eliminates feature fusion for each distinct lateral and longitudinal intention, opting instead for a straightforward feature mapping from the hidden states generated by the Intention Encoder.

Table 8.3: Ablation Test Results of GIMTP in RMSE

Horizon (Frame)	GIMTP w/o PR	GIMTP w/o DGCN	GIMTP w/o FG	GIMTP w/o FF	GIMTP
10	0.17	0.17	0.20	0.19	**0.17**
20	0.40	0.42	0.57	0.53	**0.39**
30	0.75	0.79	1.14	0.98	**0.73**
40	1.05	1.09	1.80	1.64	**1.02**
50	1.48	1.52	2.27	2.33	**1.42**

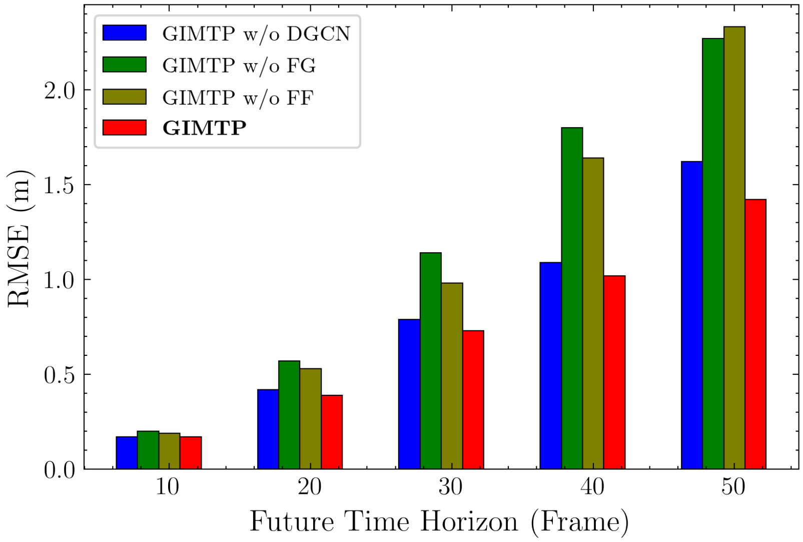


Figure 8.10: Ablation study of GIMTP.

The results, as shown in Table 8.3 and Figure 8.10, highlight several important observations. Firstly, the exclusion of the adjacency matrix influenced by potential risks (`GIMTP w/o PR`) results in a slight increase in prediction error, suggesting that potential risks contribute positively to the interaction graph's accuracy. However, the impact of this component diminishes when combined with other modules. Secondly, the omission of the DGCN module (`GIMTP w/o DGCN`) leads to a notable decline in performance, indicating that the DGCN layers are crucial for extracting dynamic graph-based information from the vehicles' motion states. Thirdly, removing the future-guided (FG) module (`GIMTP w/o FG`) causes a significant drop in performance, underscoring the importance of future-guided matrix embedding generated within the Interaction Encoder. This component effectively captures the projected future motion states of the vehicle group, which greatly influence the target vehicle's motion. Lastly, the absence of the feature fusion (FF) module (`GIMTP w/o FF`) highlights the essential role of feature fusion for each potential behavioral intention.

In conclusion, the ablation study reveals that integrating all three modules—future-guided graph embedding, feature fusion, and potential risks—substantially enhances the model's performance. Among these, the future-guided graph embedding and feature fusion demonstrate the most significant impact, indicating their critical role in optimizing the predictive model's effectiveness.

Abalation Study of RHINO Model. A detailed ablation study was conducted on the `RHINO` model to assess the impact of its individual components on prediction performance. This analysis involved systematically disabling specific components from the complete `RHINO-Gen` model to observe the resultant effects on accuracy and overall performance. The following three variants were examined:

Table 8.4: Ablation Test Results of RHINO in RMSE

Horizon (Frame)	RHINO-Gen w/o HG	RHINO-Gen w/o MM	RHINO-Gen w/o PL	RHINO-Gen
10	0.21	0.22	0.24	**0.19**
20	0.31	0.37	0.42	**0.26**
30	0.68	0.73	0.80	**0.42**
40	0.97	1.06	1.18	**0.65**
50	1.25	1.34	1.57	**0.89**

- **RHINO w/o HG**: This variant does not utilize the multiscale hypergraphs representation, relying instead on pair-wise connected graph representations within the Hypergraph Relational Encoder.

- **RHINO w/o MM**: This variant excludes multi-agent multi-modal trajectory prediction results, using only the single predicted future states for each agent as input to RHINO.

- **RHINO w/o PL**: This variant skips the Posterior Distribution Learner, directly inputting the graph embedding into the Residual Decoder.

The results, presented in Table 8.4 and Figure 8.11, reveal several critical insights. Removing various components from the RHINO model consistently leads to a degradation in performance, underscoring the importance of each component.

When the multiscale hypergraphs (RHINO w/o HG) are excluded, there is a slight increase in prediction error across the prediction horizon. This finding highlights the essential role of multiscale hypergraphs in capturing the complex, group-wise interactions among agents. Such higher-order interactions are crucial for accu-

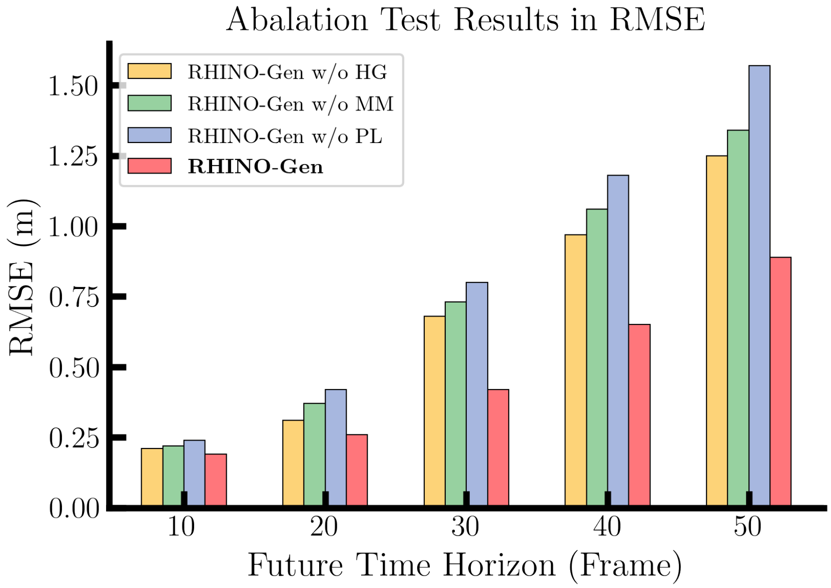


Figure 8.11: Ablation study of RHINO.

rately predicting trajectories in dynamic, multi-agent environments, as they allow the model to reason about the collective behavior of agent groups rather than just pair-wise relationships.

Excluding the multi-agent multi-modal trajectory prediction input (`RHINO w/o MM`) leads to a more significant decrease in performance. This variant's results underscore the importance of considering multiple possible future behaviors for each agent. By incorporating multi-modal predictions, the model effectively accounts for the inherent uncertainty and variability in agents' future actions, which is essential for robust and accurate trajectory prediction in real-world scenarios.

Lastly, the removal of the Posterior Distribution Learner (`RHINO w/o PL`) results in the most substantial performance decline. This component is critical for managing the stochastic nature of each agent's behavior by refining predictions through a probabilistic approach. The Posterior Distribution Learner captures the range of possible future states, thus improving the model's ability to generate accurate and reliable trajectory predictions.

In summary, the ablation study of the RHINO model reveals the relative importance of each modular component in enhancing the model's overall effectiveness. The integration of multiscale hypergraphs, multi-agent multi-modal trajectory predictions, and the Posterior Distribution Learner is vital for accurately modeling and reasoning about complex interactions among multiple agents. These components collectively contribute to significantly improved performance metrics, demonstrating their essential roles in the RHINO model's architecture.

Chapter 9

Conclusion

This thesis aims to develop an integrated learning-based framework for motion prediction and planning in connected and automated vehicles (CAVs) operating in dynamic traffic environments. The primary objective is to address the complexities associated with vehicle interactions, driving behaviors, and interaction relational reasoning, thereby providing robust and reliable predictions and plans that enhance the safety and efficiency of autonomous driving systems.

The core contributions of this thesis are embodied in the development of the `GIMTP`, `GIRAFFE`, and `RHINO` models. `GIMTP` (Graph-based Interaction-aware Multi-agent Multi-modal Trajectory Prediction) introduces a sophisticated framework that deeply investigates vehicle interactions, offering multiple potential predictions and probabilistically estimating driving behavioral intentions. A dynamic adjacency matrix captures comprehensive vehicle interactions by considering neighborhood, distance, and potential risk factors. The implementation of the Diffusion

Graph Convolutional Network (DGCN) structure allows for the encapsulation of both spatial and temporal vehicle interactions. This model integrates not only the historical motion states of the vehicle group but also the inferred future motion states, providing additional correlations and potential future interactions. Feature fusion is employed to efficiently combine historical and future embeddings for collective intention recognition and trajectory prediction, facilitating more precise trajectory generation based on latent variables representing multi-modal behaviors. The model anticipates both longitudinal and lateral driving behaviors in a multi-modal manner, associating potential future trajectories with corresponding probabilities.

Building upon GIMTP, the enhanced GIRAFFE model further improves prediction accuracy by refining the multi-agent prediction function and optimizing computational efficiency. This model's architecture incorporates intention prediction and multi-modal decoding, enabling the generation of multiple future trajectory distributions with associated probabilities.

The RHINO (Hypergraph-based Interaction Relational Reasoning Motion Generation and Planning) framework represented a significant leap forward by introducing hypergraph representations to model high-dimensional and group-wise social interactions between various modalities of behaviors and motion states of multiple agents. This model comprised several critical modules: the Hypergraph Relational Encoder, Posterior Distribution Learner, Residual Decoder, and Neural Planner and Refinement. RHINO employs representation learning to enable explicit interaction relational reasoning. By inferring the multiscale hypergraph topology, our models are capable of providing social behavior-inspired automated driving. This involves considering future relations and interactions and learning the posterior distribution to handle the stochasticity of social behavior for each agent. The models generate plausible motion planning in a generative manner and offer planned trajectory refinements that prioritize safety, efficiency, and comfort, ultimately enabling human-like automated driving. We validated the proposed model using real-world trajectory datasets, confirming its effectiveness and applicability in practical scenarios.

The integration of GIMTP, GIRAFFE, and RHINO into a cohesive framework provides a comprehensive approach to trajectory prediction and motion planning. This overall framework excels in capturing the intricate interactions between multiple agents, modeling the stochasticity of driving behaviors, and generating refined, feasible trajectories that account for safety, efficiency, and comfort. By leveraging advanced graph-based and hypergraph-based representations, the framework ensures robust performance in dynamic and complex traffic environments. The modular nature of the framework allows for scalable and flexible adaptations to various traffic scenarios, thereby significantly advancing the state-of-the-art in automated driving and multi-agent trajectory prediction.

While this thesis presents significant advancements, there are notable limitations that suggest avenues for future research. The formulation of large-scale hypergraphs is not computationally efficient, indicating the need for enhanced graph and hypergraph structure learning. Current models struggle with effectively capturing temporal correlations between interactions; thus, future work should focus on developing dynamic hypergraph representations. Accurately modeling the temporal evolution of interactions between agents is crucial for predicting future behaviors in dynamic environments. This involves creating representations that can adapt and change as the interactions among vehicles evolve over time, ensuring that the models remain responsive to the fluid nature of traffic scenarios.

There is a notable gap between planned trajectories and actual vehicle control inputs, which can be addressed by developing differentiable cost functions for neural planners to improve trajectory planning and refinement. This discrepancy arises because the planned trajectories often do not consider the detailed dynamics and control limitations of the vehicles. For instance, a planned trajectory might suggest a sharp turn or sudden acceleration that is not feasible given the vehicle's mechanical constraints and current speed. To bridge this gap, future models should incorporate differentiable cost functions that consider these physical constraints during the planning process. By doing so, the neural planner can generate more realistic and feasible trajectories that the vehicle can follow accurately. Additionally, incorporating feedback from the actual control inputs back into the planning algorithm can help refine and adjust the planned trajectories in real-time, leading to more accurate and reliable motion planning.

The estimation of surrounding vehicles often lacks contextual information, necessitating the incorporation of vehicle-road, vehicle-environment, and road-road interactions. Finally, the input data often lacks realistic assumptions, such as data collection delays and errors, underscoring the importance of incorporating more realistic assumptions and context information into future models.

These future improvements will enhance the robustness and applicability of hypergraph-based models, making them more efficient and capable of handling the complexities of real-world driving scenarios.

Bibliography

[1] Bin Ran, Yang Cheng, Shen Li, Fan Ding, Jing Jin, Xiaoxuan Chen, and Zhen Zhang. Connected automated vehicle highway systems and methods, August 13 2019. US Patent 10,380,886.

[2] Morteza Taiebat, Austin L Brown, Hannah R Safford, Shen Qu, and Ming Xu. A review on energy, environmental, and sustainability implications of connected and automated vehicles. *Environmental science & technology*, 52 (20):11449–11465, 2018.

[3] Steven E Shladover. Connected and automated vehicle systems: Introduction and overview. *Journal of Intelligent Transportation Systems*, 22(3):190–200, 2018.

[4] Hongjie Liu, Keshu Wu, Sicheng Fu, Haotian Shi, and Hongzhe Xu. Predictive analysis of vehicular lane changes: An integrated lstm approach. *Applied Sciences*, 13(18):10157, 2023.

[5] Kunsong Shi, Yuankai Wu, Haotian Shi, Yang Zhou, and Bin Ran. An integrated car-following and lane changing vehicle trajectory prediction algorithm based on a deep neural network. *Physica A: Statistical Mechanics and its Applications*, 599:127303, 2022.

[6] Jinxin Liu, Yugong Luo, Hui Xiong, Tinghan Wang, Heye Huang, and Zhihua Zhong. An integrated approach to probabilistic vehicle trajectory prediction via driver characteristic and intention estimation. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 3526–3532. IEEE, 2019.

[7] Haotian Shi, Danjue Chen, Nan Zheng, Xin Wang, Yang Zhou, and Bin Ran. A deep reinforcement learning based distributed control strategy for connected automated vehicles in mixed traffic platoon. *Transportation Research Part C: Emerging Technologies*, 148:104019, 2023.

[8] Jacopo Guanetti, Yeojun Kim, and Francesco Borrelli. Control of connected and automated vehicles: State of the art and future challenges. *Annual reviews in control*, 45:18–40, 2018.

[9] Haotian Shi, Yang Zhou, Keshu Wu, Xin Wang, Yangxin Lin, and Bin Ran. Connected automated vehicle cooperative control with a deep reinforcement learning approach in a mixed traffic environment. *Transportation Research Part C: Emerging Technologies*, 133:103421, 2021.

[10] Haotian Shi, Yang Zhou, Keshu Wu, Sikai Chen, Bin Ran, and Qinghui Nie. Physics-informed deep reinforcement learning-based integrated two-dimensional car-following control strategy for connected automated vehicles. *Knowledge-Based Systems*, 269:110485, 2023.

[11] Vinicius Trentin, Antonio Artuñedo, Jorge Godoy, and Jorge Villagra. Multimodal interaction-aware motion prediction at unsignalized intersections. *IEEE Transactions on Intelligent Vehicles*, 2023.

[12] Phillip Karle, Maximilian Geisslinger, Johannes Betz, and Markus Lienkamp. Scenario understanding and motion prediction for autonomous vehicles—review and comparison. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):16962–16982, 2022.

[13] Francesco Marchetti, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Multiple trajectory prediction of moving agents with memory augmented networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6688–6702, 2020.

[14] Shuo Feng, Xintao Yan, Haowei Sun, Yiheng Feng, and Henry X Liu. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. *Nature communications*, 12(1):748, 2021.

[15] Siddhesh Khandelwal, William Qi, Jagjeet Singh, Andrew Hartnett, and Deva Ramanan. What-if motion prediction for autonomous driving. *arXiv preprint arXiv:2008.10587*, 2020.

[16] Yanjun Huang, Jiatong Du, Ziru Yang, Zewei Zhou, Lin Zhang, and Hong Chen. A survey on trajectory-prediction methods for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 7(3):652–674, 2022.

[17] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

[18] Xin Li, Xiaowen Ying, and Mooi Choo Chuah. Grip: Graph-based interaction-aware trajectory prediction. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 3960–3966. IEEE, 2019.

[19] Nachiket Deo and Mohan M Trivedi. Multi-modal trajectory prediction of surrounding vehicles with maneuver based lstms. In *2018 IEEE intelligent vehicles symposium (IV)*, pages 1179–1184. IEEE, 2018.

[20] Chenxin Xu, Maosen Li, Zhenyang Ni, Ya Zhang, and Siheng Chen. Groupnet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6507, 2022.

[21] Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. *Advances in neural information processing systems*, 33:19783–19794, 2020.

[22] Zhichao Han, Olga Fink, and David S Kammer. Collective relational inference for learning heterogeneous interactions. *Nature Communications*, 15(1): 3191, 2024.

[23] Djamel Eddine Benrachou, Sebastien Glaser, Mohammed Elhenawy, and Andry Rakotonirainy. Use of social interaction and intention to improve motion prediction within automated vehicle framework: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):22807–22837, 2022.

[24] Hua Wang, Qiang Meng, Shukai Chen, and Xiaoning Zhang. Competitive and cooperative behaviour analysis of connected and autonomous vehicles across unsignalised intersections: A game-theoretic approach. *Transportation research part B: methodological*, 149:322–346, 2021.

[25] Rui Zhou, Hongyu Zhou, Huidong Gao, Masayoshi Tomizuka, Jiachen Li, and Zhuo Xu. Grouptron: Dynamic multi-scale graph convolutional networks for group-aware dense crowd trajectory forecasting. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 805–811. IEEE, 2022.

[26] Jiachen Li, Chuanbo Hua, Hengbo Ma, Jinkyoo Park, Victoria Dax, and Mykel J Kochenderfer. Multi-agent dynamic relational reasoning for social robot navigation. *arXiv preprint arXiv:2401.12275*, 2024.

[27] Zhang Zhang, Yifeng Zeng, Wenhui Jiang, Yinghui Pan, and Jing Tang. Intention recognition for multiple agents. *Information Sciences*, 628:360–376, 2023.

[28] SAE International. Taxonomy and definitions for terms related to cooperative driving automation for on-road motor vehicles j3216_202107, 2021. URL `https://www.sae.org/standards/content/j3216_202107/`.

[29] Keshu Wu, Yang Zhou, Haotian Shi, Xiaopeng Li, and Bin Ran. Graph-based interaction-aware multimodal 2d vehicle trajectory prediction using diffu-

sion graph convolutional networks. *IEEE Transactions on Intelligent Vehicles*, 9(2):3630–3643, 2024. doi: 10.1109/TIV.2023.3341071.

[30] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11525–11533, 2020.

[31] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):3316–3333, 2021.

[32] Dennis C Rapaport. *The art of molecular dynamics simulation*. Cambridge university press, 2004.

[33] Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54 (2):875–891, 2011.

[34] Adam Houenou, Philippe Bonnifait, Véronique Cherfaoui, and Wen Yao. Vehicle trajectory prediction based on motion model and maneuver recognition. In *2013 IEEE/RSJ international conference on intelligent robots and systems*, pages 4363–4369. IEEE, 2013.

[35] Chiu-Feng Lin, A Galip Ulsoy, and David J LeBlanc. Vehicle dynamics and external disturbance estimation for vehicle path prediction. *IEEE Transactions on Control Systems Technology*, 8(3):508–518, 2000.

[36] ByeoungDo Kim, Chang Mook Kang, Jaekyum Kim, Seung Hi Lee, Chung Choo Chung, and Jun Won Choi. Probabilistic vehicle trajectory prediction over occupancy grid map via recurrent neural network. In *2017 IEEE 20Th international conference on intelligent transportation systems (ITSC)*, pages 399–404. IEEE, 2017.

[37] Georges S Aoude, Vishnu R Desaraju, Lauren H Stephens, and Jonathan P How. Driver behavior classification at intersections and validation on large naturalistic data set. *IEEE Transactions on Intelligent Transportation Systems*, 13(2):724–736, 2012.

[38] Dizan Vasquez, Thierry Fraichard, and Christian Laugier. Incremental learning of statistical motion patterns with growing hidden markov models. *IEEE Transactions on intelligent transportation systems*, 10(3):403–416, 2009.

[39] Sebastian Thrun and Michael Montemerlo. The graph slam algorithm with applications to large-scale mapping of urban structures. *The International Journal of Robotics Research*, 25(5-6):403–429, 2006.

[40] Kaouther Messaoud, Itheri Yahiaoui, Anne Verroust-Blondet, and Fawzi Nashashibi. Attention based vehicle trajectory prediction. *IEEE Transactions on Intelligent Vehicles*, 6(1):175–185, 2020.

[41] Derek J. Phillips, Tim A. Wheeler, and Mykel J. Kochenderfer. Generalizable intention prediction of human drivers at intersections. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 1665–1670, 2017. doi: 10.1109/IVS.2017.7995948.

[42] Nachiket Deo and Mohan M Trivedi. Convolutional social pooling for vehicle trajectory prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1468–1476, 2018.

[43] Daniela Ridel, Nachiket Deo, Denis Wolf, and Mohan Trivedi. Scene compliant trajectory forecast with agent-centric spatio-temporal grids. *IEEE Robotics and Automation Letters*, 5(2):2816–2823, 2020.

[44] Henggang Cui, Vladan Radosavljevic, Fang-Chieh Chou, Tsung-Han Lin, Thi Nguyen, Tzu-Kuo Huang, Jeff Schneider, and Nemanja Djuric. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *2019 international conference on robotics and automation (icra)*, pages 2090–2096. IEEE, 2019.

[45] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[46] Kaouther Messaoud, Itheri Yahiaoui, Anne Verroust-Blondet, and Fawzi Nashashibi. Non-local social pooling for vehicle trajectory prediction. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 975–980. IEEE, 2019.

[47] Yu Wang, Shengjie Zhao, Rongqing Zhang, Xiang Cheng, and Liuqing Yang. Multi-vehicle collaborative learning for trajectory prediction with spatio-temporal tensor fusion. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):236–248, 2020.

[48] Tianyang Zhao, Yifei Xu, Mathew Monfort, Wongun Choi, Chris Baker, Yibiao Zhao, Yizhou Wang, and Ying Nian Wu. Multi-agent tensor fusion for contextual trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12126–12134, 2019.

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[50] Xiaobo Chen, Huanjia Zhang, Feng Zhao, Yu Hu, Chenkai Tan, and Jian Yang. Intention-aware vehicle trajectory prediction based on spatial-temporal dynamic attention network for internet of vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):19471–19483, 2022.

[51] Kunpeng Zhang, Liang Zhao, Chengxiang Dong, Lan Wu, and Liang Zheng. Ai-tp: Attention-based interaction-aware trajectory prediction for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 8(1):73–83, 2022.

[52] Hayoung Kim, Dongchan Kim, Gihoon Kim, Jeongmin Cho, and Kunsoo Huh. Multi-head attention based probabilistic vehicle trajectory prediction. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, pages 1720–1725. IEEE, 2020.

[53] Kaouther Messaoud, Nachiket Deo, Mohan M Trivedi, and Fawzi Nashashibi. Trajectory prediction for autonomous driving based on multi-head attention with joint agent-map representation. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 165–170. IEEE, 2021.

[54] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[55] Hao Zhou, Dongchun Ren, Huaxia Xia, Mingyu Fan, Xu Yang, and Hai Huang. Ast-gnn: An attention-based spatio-temporal graph neural network for interaction-aware pedestrian trajectory prediction. *Neurocomputing*, 445: 298–308, 2021.

[56] Yingfan Huang, Huikun Bi, Zhaoxin Li, Tianlu Mao, and Zhaoqi Wang. Stgat: Modeling spatial-temporal interactions for human trajectory prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6272–6281, 2019.

[57] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, pages 683–700. Springer, 2020.

[58] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Mo Zhou, Zhenxing Niu, and Gang Hua. Sgcn: Sparse graph convolution network for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8994–9003, 2021.

[59] Zihao Sheng, Yunwen Xu, Shibei Xue, and Dewei Li. Graph-based spatial-temporal convolutional network for vehicle trajectory prediction in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23 (10):17654–17665, 2022.

[60] Daehee Park, Hobin Ryu, Yunseo Yang, Jegyeong Cho, Jiwon Kim, and Kuk-Jin Yoon. Leveraging future relationship reasoning for vehicle trajectory prediction. *arXiv preprint arXiv:2305.14715*, 2023.

[61] Srikanth Malla, Chiho Choi, and Behzad Dariush. Social-stage: Spatio-temporal multi-modal future trajectory forecast. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13938–13944. IEEE, 2021.

[62] Javad Amirian, Jean-Bernard Hayet, and Julien Pettré. Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[63] Junru Gu, Chen Sun, and Hang Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15303–15312, 2021.

[64] Wenyuan Zeng, Ming Liang, Renjie Liao, and Raquel Urtasun. Lanercnn: Distributed representations for graph-centric motion forecasting. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 532–539. IEEE, 2021.

[65] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35:6531–6543, 2022.

[66] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Mtr++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[67] Chenxu Luo, Lin Sun, Dariush Dabiri, and Alan Yuille. Probabilistic multi-modal trajectory prediction with lane attention for autonomous vehicles.

In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2370–2376. IEEE, 2020.

[68] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[69] Xidong Feng, Zhepeng Cen, Jianming Hu, and Yi Zhang. Vehicle trajectory prediction using intention-based conditional variational autoencoder. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 3514–3519. IEEE, 2019.

[70] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[71] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2255–2264, 2018.

[72] Haoran Song, Di Luan, Wenchao Ding, Michael Y Wang, and Qifeng Chen. Learning to predict vehicle trajectories with model-based planning. In *Conference on Robot Learning*, pages 1035–1045. PMLR, 2022.

[73] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1349–1358, 2019.

[74] Jiachen Li, Hengbo Ma, and Masayoshi Tomizuka. Conditional generative neural system for probabilistic trajectory prediction. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6150–6156. IEEE, 2019.

[75] Sergio Casas, Cole Gulino, Simon Suo, Katie Luo, Renjie Liao, and Raquel Urtasun. Implicit latent variable model for scene-consistent motion forecasting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 624–641. Springer, 2020.

[76] Seong Hyeon Park, ByeongDo Kim, Chang Mook Kang, Chung Choo Chung, and Jun Won Choi. Sequence-to-sequence prediction of vehicle trajectory via lstm encoder-decoder architecture. In *2018 IEEE intelligent vehicles symposium (IV)*, pages 1672–1678. IEEE, 2018.

[77] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000. doi: 10.1126/science.290.5500.2323. URL `https://www.science.org/doi/abs/10.1126/science.290.5500.2323`.

[78] Joshua B. Tenenbaum, Vin de Silva, and John C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290 (5500):2319–2323, 2000. doi: 10.1126/science.290.5500.2319. URL `https://www.science.org/doi/abs/10.1126/science.290.5500.2319`.

[79] Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems. In *International conference on machine learning*, pages 2688–2697. PMLR, 2018.

[80] Chenxin Xu, Yuxi Wei, Bohan Tang, Sheng Yin, Ya Zhang, Siheng Chen, and Yanfeng Wang. Dynamic-group-aware networks for multi-agent trajectory prediction with relational reasoning. *Neural Networks*, 170:564–577, 2024.

[81] Peng Hang, Chen Lv, Chao Huang, Jiacheng Cai, Zhongxu Hu, and Yang Xing. An integrated framework of decision making and motion planning for autonomous vehicles considering social behaviors. *IEEE transactions on vehicular technology*, 69(12):14458–14469, 2020.

[82] Zhiyu Huang, Jingda Wu, and Chen Lv. Driving behavior modeling using naturalistic human driving data with inverse reinforcement learning. *IEEE transactions on intelligent transportation systems*, 23(8):10239–10251, 2021.

[83] Hao Zhou, Jorge Laval, Anye Zhou, Yu Wang, Wenchao Wu, Zhu Qing, and Srinivas Peeta. Review of learning-based longitudinal motion planning for autonomous vehicles: research gaps between self-driving and traffic congestion. *Transportation research record*, 2676(1):324–341, 2022.

[84] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023.

[85] Zhiyu Huang, Haochen Liu, and Chen Lv. Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3903–3913, 2023.

[86] Haochen Liu, Zhiyu Huang, and Chen Lv. Occupancy prediction-guided neural planner for autonomous driving. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pages 4859–4865. IEEE, 2023.

[87] Haoran Song, Wenchao Ding, Yuxuan Chen, Shaojie Shen, Michael Yu Wang, and Qifeng Chen. Pip: Planning-informed trajectory prediction for autonomous driving. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 598–614. Springer, 2020.

[88] Zhiyu Huang, Haochen Liu, Jingda Wu, and Chen Lv. Differentiable integrated motion prediction and planning with learnable cost function for autonomous driving. *IEEE transactions on neural networks and learning systems*, 2023.

[89] Jose Luis Vazquez Espinoza, Alexander Liniger, Wilko Schwarting, Daniela Rus, and Luc Van Gool. Deep interactive motion prediction and planning: Playing games with motion prediction models. In *Learning for Dynamics and Control Conference*, pages 1006–1019. PMLR, 2022.

[90] Bin Ran, Yang Cheng, Shawn Leight, and Steven Parker. Development of an integrated transportation system of connected automated vehicles and highways. *ITE Journal*, 89(11), 2019.

[91] Bin Ran, Yuan Zheng, Kaijie Luo, Haozhan Ma, Yikang Rui, Linheng Li, Xiaolong Li, Jinling Hu, and Yanming Hu. Architecture design of a vehicle–road-cloud collaborative automated driving system. *Urban Lifeline*, 1(1):9, 2023.

[92] Bin Ran, Yang Cheng, Shen Li, Hanchu Li, and Steven Parker. Classification of roadway infrastructure and collaborative automated driving system. *SAE International Journal of Connected and Automated Vehicles*, 6(12-06-04-0026), 2023.

[93] 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*, 1(2):70–76, 2017.

[94] Laurens Hobert, Andreas Festag, Ignacio Llatser, Luciano Altomare, Filippo Visintainer, and Andras Kovacs. Enhancements of v2x communication in support of cooperative autonomous driving. *IEEE communications magazine*, 53(12):64–70, 2015.

[95] Pei Li, Keshu Wu, Yang Cheng, Steven T Parker, and David A Noyce. How does c-v2x perform in urban environments? results from real-world experiments on urban arterials. *IEEE Transactions on Intelligent Vehicles*, 2023.

[96] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vi-

sion transformer. In *European conference on computer vision*, pages 107–124. Springer, 2022.

[97] SAE International. V2x communications message set dictionary j2735_202309, 2023. URL `https://www.sae.org/standards/content/j2735_202309/`.

[98] Igal Bilik, Oren Longman, Shahar Villeval, and Joseph Tabrikian. The rise of radar for autonomous vehicles: Signal processing solutions and future research directions. *IEEE signal processing Magazine*, 36(5):20–31, 2019.

[99] Juergen Dickmann, Jens Klappstein, Markus Hahn, Nils Appenrodt, Hans-Ludwig Bloecher, Klaudius Werber, and Alfons Sailer. Automotive radar the key technology for autonomous driving: From detection and ranging to environmental understanding. In *2016 IEEE Radar Conference (RadarConf)*, pages 1–6. IEEE, 2016.

[100] Jessica Van Brummelen, Marie O'brien, Dominique Gruyer, and Homayoun Najjaran. Autonomous vehicle perception: The technology of today and tomorrow. *Transportation research part C: emerging technologies*, 89:384–406, 2018.

[101] Jelena Kocić, Nenad Jovičić, and Vujo Drndarević. Sensors and sensor fusion in autonomous vehicles. In *2018 26th Telecommunications Forum (TELFOR)*, pages 420–425. IEEE, 2018.

[102] Xiangmo Zhao, Pengpeng Sun, Zhigang Xu, Haigen Min, and Hongkai Yu. Fusion of 3d lidar and camera data for object detection in autonomous vehicle applications. *IEEE Sensors Journal*, 20(9):4901–4913, 2020.

[103] Santiago Royo and Maria Ballesta-Garcia. An overview of lidar imaging systems for autonomous vehicles. *Applied sciences*, 9(19):4093, 2019.

[104] You Li and Javier Ibanez-Guzman. Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. *IEEE Signal Processing Magazine*, 37(4):50–61, 2020.

[105] Shan Su, Cheng Peng, Jianbo Shi, and Chiho Choi. Potential field: Interpretable and unified representation for trajectory prediction. *arXiv preprint arXiv:1911.07414*, 2019.

[106] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995.

[107] Yadollah Rasekhipour, Amir Khajepour, Shih-Ken Chen, and Bakhtiar Litkouhi. A potential field-based model predictive path-planning controller for autonomous road vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 18(5):1255–1267, 2016.

[108] Neel P Bhatt, Amir Khajepour, and Ehsan Hashemi. Mpc-pf: Social interaction aware trajectory prediction of dynamic objects for autonomous driving using potential fields. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9837–9844. IEEE, 2022.

[109] Xunjia Zheng, Huilan Li, Hui Liu, Xing Chen, and Tianhong Luo. A modeling method of driving risk assessment based on vehicle trajectory prediction. In *2022 IEEE 2nd International Conference on Digital Twins and Parallel Intelligence (DTPI)*, pages 1–4. IEEE, 2022.

[110] Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.

[111] Yuankai Wu, Dingyi Zhuang, Aurelie Labbe, and Lijun Sun. Inductive graph neural networks for spatiotemporal kriging. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4478–4485, 2021.

[112] Artidoro Pagnoni, Kevin Liu, and Shangyan Li. Conditional variational autoencoder for neural machine translation. *arXiv preprint arXiv:1812.04405*, 2018.

[113] J. Colyar and J. Halkias. U.s. highway 101 dataset, 2007.

[114] J. Colyar and J. Halkias. U.s. highway 80 dataset, 2006.

[115] Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st international conference on intelligent transportation systems (ITSC)*, pages 2118–2125. IEEE, 2018.