HeLoFusion: An Efficient and Scalable Encoder for Modeling Heterogeneous and Multi-Scale Interactions in Trajectory Prediction

Bingqing Wei* **Lianmin Chen*†** **Zhongyu Xia** **Yongtao Wang‡**
Wangxuan Institute of Computer Technology, Peking University
bingqing.wei@stu.pku.edu.cn lianminchen@hust.edu.cn
{xiazhongyu,wyt}@pku.edu.cn

Abstract

Multi-agent trajectory prediction in autonomous driving requires a comprehensive understanding of complex social dynamics. Existing methods, however, often struggle to capture the full richness of these dynamics, particularly the co-existence of multi-scale interactions and the diverse behaviors of heterogeneous agents. To address these challenges, this paper introduces HeLoFusion, an efficient and scalable encoder for modeling heterogeneous and multi-scale agent interactions. Instead of relying on global context, HeLoFusion constructs local, multi-scale graphs centered on each agent, allowing it to effectively model both direct pairwise dependencies and complex group-wise interactions (*e.g.*, platooning vehicles or pedestrian crowds). Furthermore, HeLoFusion tackles the critical challenge of agent heterogeneity through an aggregation-decomposition message-passing scheme and type-specific feature networks, enabling it to learn nuanced, type-dependent interaction patterns. This locality-focused approach enables a principled representation of multi-level social context, yielding powerful and expressive agent embeddings. On the challenging Waymo Open Motion Dataset, HeLoFusion achieves state-of-the-art performance, setting new benchmarks for key metrics including Soft mAP and minADE. Our work demonstrates that a locality-grounded architecture, which explicitly models multi-scale and heterogeneous interactions, is a highly effective strategy for advancing motion forecasting.

1 Introduction

Trajectory prediction, forecasting the future movements of traffic participants, is crucial for autonomous driving. This task is challenging because it involves understanding complex, multi-scale social dynamics and the diverse behaviors of heterogeneous agents (*e.g.*, vehicles vs. pedestrians). While many deep learning methods have been proposed, they often fall short in one of two ways. Early approaches, such as social pooling (Gupta et al. [2018], Deo and Trivedi [2018], Hasan et al. [2021]), are efficient but model interactions too simplistically. Conversely, more recent methods that rely on global context, such as those using global attention (Liu et al. [2021], Ngiam et al. [2021], Shi et al. [2022]) or dense graph neural networks (GNNs) (Shi et al. [2021], Xu et al. [2022]), attempt to capture all possible dependencies. However, this global view can be suboptimal, as it may dilute the influence of critical nearby agents with information from distant, irrelevant ones, while also incurring substantial computational costs. Furthermore, explicitly modeling interactions between

*These authors contributed equally to this work.

†Lianmin Chen completed this work during his internship at the Wangxuan Institute of Computer Technology.

‡Corresponding author: Yongtao Wang (wyt@pku.edu.cn).

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

heterogeneous agent types often leads to a combinatorial explosion in model parameters, forcing many models to treat all agents homogeneously.

To address these fundamental modeling challenges, we propose the **He**terogeneous **Lo**cal Context **Fusion** Network (**HeLoFusion**), a novel encoder designed to capture the rich structure of social interactions in a more principled manner. Our core insight is that social dynamics are inherently multi-level and type-dependent. HeLoFusion explicitly models this by constructing localized, multi-scale graphs around each agent. This allows it to capture both direct pairwise dependencies and complex group-wise interactions (*e.g.*, platooning vehicles or pedestrian crowds), which are often overlooked by methods with a flat interaction structure. Furthermore, HeLoFusion tackles the critical challenge of agent heterogeneity through an aggregation-decomposition message-passing scheme and type-specific feature networks, making it learn nuanced, type-dependent interaction patterns without a prohibitive increase in parameters. In the end, by grounding these sophisticated modeling capabilities in the principle of spatial locality, our approach not only aligns with the real-world observation that an agent's behavior is primarily influenced by its immediate surroundings but also yields a highly efficient and scalable architecture.

Our key contributions are summarized as follows:

1. We propose a novel graph-based module that efficiently models both **pairwise and group-wise interactions** among traffic participants by constructing multi-scale local graphs.

2. We jointly exploit **agent heterogeneity** and **spatial locality** throughout the encoding process: category-specific networks and an aggregation–decomposition message-passing scheme operate on local neighborhoods to capture type-dependent interaction dynamics in a computationally efficient manner.

3. Extensive experiments on the Waymo Open Motion Dataset (WOMD) demonstrate that HeLoFusion achieves **state-of-the-art performance** among comparable methods, with clear gains in Soft mAP, displacement errors, and overlap rate over strong baselines.

2 Related Work

The development of large-scale datasets (Ettinger et al. [2021]) has spurred significant progress in motion prediction, with deep learning emerging as the standard approach. Early works often visualized the scene as a top-down rasterized image, leveraging convolutional networks to forecast future states (Casas et al. [2020], Luo et al. [2021], Zeng et al. [2021], Konev et al. [2022]). More recently, the field has shifted towards vectorized representations, where Transformer-based architectures (Jiang et al. [2023], Seff et al. [2023], Ngiam et al. [2021], Nayakanti et al. [2023], Varadarajan et al. [2022], Shi et al. [2022], Zhou et al. [2023]) and Graph Neural Networks (GNNs) (Jia et al. [2023], Mo et al. [2022], Cui et al. [2022], Salzmann et al. [2020]) are widely used to model complex social interactions. These models typically aim to capture future uncertainty by predicting either a dense occupancy grid or a sparse set of multi-modal future trajectories. State-of-the-art methods, such as MTR++ (Shi et al. [2024]) and its successors (Lin et al. [2024], Sun et al. [2024b], Liu et al. [2024], Sun et al. [2024a], Yan et al. [2025]), have achieved remarkable performance by capturing a global context of the scene. However, this global view can be computationally intensive and may dilute the influence of critical nearby agents. In contrast, our work is built on the principle of spatial locality, focusing on modeling structured, multi-scale interactions within local neighborhoods while explicitly addressing the heterogeneity of different agent types.

3 Method

Given historical trajectories $\mathcal{O} = \{\mathbf{X}_i^- \mid i \in \mathcal{A}\}$ for all agents $\mathcal{A}$ and map information $\mathcal{M}$, our goal is to learn a rich embedding for each agent to predict its future trajectory $\hat{\mathbf{X}}_a^+$. HeLoFusion achieves this through a multi-stage encoding process that first extracts individual motion features, then models heterogeneous, multi-scale interactions, and finally fuses local scene context, as shown in Figure 1.

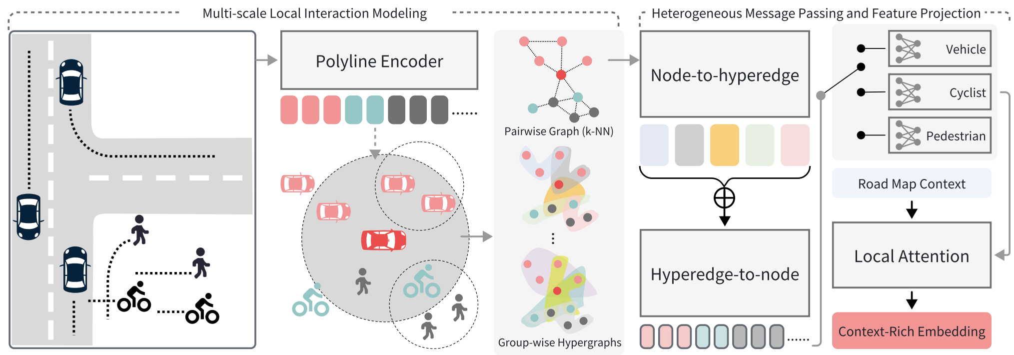

Figure 1: **Overview of the HeLoFusion architecture.** The proposed encoder builds local multi-scale graphs (pairwise and group-wise) to model social interactions. It then employs a heterogeneous message-passing scheme and type-specific projection to capture the diverse dynamics of agents. Finally, a local attention module fuses the agent representations with the road map context to produce a context-rich embedding.

3.1 Capturing Pairwise and Group-wise Interactions with Local Graphs

A core contribution of our work is the efficient modeling of complex interactions. Inspired by the spatial locality of real-world traffic scenarios, we discard expensive global graph operations and instead construct local graphs centered on each agent of interest.

Local Graph Construction. For each target agent, we identify its K nearest neighbors. To model pairwise interactions, we construct a k-NN graph that connects these neighbors. To capture more complex group-wise interactions (*e.g.*, a platoon of vehicles or a crowd of pedestrians), we construct local hypergraphs. Each hyperedge connects a neighboring agent to its own small group of nearby agents, effectively representing a local cluster. By varying the size of these groups, we create a series of graphs that capture interactions at multiple scales. This multi-scale, localized structure provides a rich yet computationally tractable foundation for modeling social dynamics.

3.2 Modeling Agent and Interaction Heterogeneity

Traffic participants are not homogeneous; vehicles, pedestrians, and cyclists behave and interact differently. HeLoFusion explicitly models this heterogeneity at multiple stages.

Heterogeneous Message Passing. To handle interactions between different agent types without a combinatorial explosion of parameters (*e.g.*, separate edge types for car-pedestrian, car-cyclist, etc.), we use an aggregation-decomposition message passing scheme inspired by Xu et al. [2022]. Our model first aggregates feature information from all nodes connected by an edge into a single representation. Then, a small, shared MLP dynamically decomposes this aggregated influence into category-specific messages based on the types of the participating agents. This allows interaction effects to be flexibly tailored to agent categories in a scalable manner.

Heterogeneous Feature Projection. To further reinforce type-specific behaviors, agent features are processed through a bank of category-specific MLPs. Each agent's embedding is routed through the MLP corresponding to its type (*e.g.*, vehicle, pedestrian). This ensures that the feature representations are specialized for the distinct motion patterns and constraints of each category before being fused with scene context.

3.3 Overall Architecture and Feature Fusion

The principles of localized, heterogeneous interaction modeling are embedded within a three-stage encoder architecture.

Motion Encoding. We first process each agent's historical trajectory and all map polylines using a PointNet-style polyline encoder. This module treats a trajectory as an unordered set of points, making

Table 1: **Performance comparison on WOMD Motion Leaderboard.** This table shows results for motion predictors without model ensemble or using extra data such as LIDAR on the `Test` set (except for mAP-`Val`). The best and second best results are **bolded** and underlined, respectively.

Method	minADE ↓	minFDE ↓	MR (%) ↓	OR (%) ↓	**mAP-`Val` (%) ↑**	**mAP (%) ↑**	**Soft mAP (%) ↑**
MTR++ (Shi et al. [2024])	0.5906	1.1939	12.98	12.81	43.51	43.29	44.10
EDA (Lin et al. [2024])	0.5718	1.1702	11.69	12.66	43.53	44.87	45.96
ControlMTR (Sun et al. [2024b])	0.5897	1.1916	12.82	<u>12.59</u>	44.13	44.14	45.72
BehaveOcc	0.5723	1.1668	11.76	12.78	-	45.66	46.78
BeTopNet (Liu et al. [2024])	0.5716	1.1668	11.83	12.72	44.16	45.87	46.98
RMP-YOLO (Sun et al. [2024a])	0.5737	1.1697	**11.60**	12.66	-	45.23	46.73
TrajFlow (Yan et al. [2025])	<u>0.5714</u>	<u>1.1667</u>	<u>11.62</u>	12.72	<u>45.39</u>	<u>46.04</u>	<u>47.10</u>
HeLoFusion (Ours)	**0.5690**	**1.1596**	11.83	**12.43**	**46.37**	**46.24**	**47.32**

it robust to irregular sampling and permutation-invariant. It produces a compact feature vector for each agent and map element, with agent type embeddings included to provide initial heterogeneous signals.

Interaction Modeling. The motion features are then fed into the local multi-scale graph network described in Section 3.1. The heterogeneous message passing scheme (Section 3.2) is applied to produce socially-aware agent embeddings that incorporate both pairwise and group-wise dependencies.

Context Fusion. Finally, the interaction-aware embeddings are refined using a heterogeneous local attention module. After being projected by category-specific MLPs (Section 3.2), each agent attends to its nearby neighbors and map elements. This localized attention mechanism efficiently integrates dynamic agent information with static map constraints, producing the final context-rich embedding for the downstream prediction decoder.

4 Experiments

4.1 Experimental Setup

We evaluate HeLoFusion on the large-scale Waymo Open Motion Dataset (WOMD) (Ettinger et al. [2021]), a standard benchmark for multi-agent trajectory forecasting in autonomous driving. Following the official protocol, we report key metrics including mean Average Precision (mAP), Soft mAP, minimum Average Displacement Error (minADE), minimum Final Displacement Error (minFDE), Miss Rate (MR), and Overlap Rate (OR), all evaluated with $K = 6$ predictions per agent. Our model is designed as a modular encoder; for a fair and direct comparison, we replace the encoder of the BeTopNet framework (Liu et al. [2024]) with HeLoFusion, while maintaining the identical decoder architecture. We compare our method against other top-performing, publicly reported methods on the WOMD leaderboard that do not use extra data such as LIDAR or ensemble techniques. More implementation details can be found in Appendix A.

4.2 Results

As shown in Table 1, HeLoFusion achieves state-of-the-art performance on the WOMD test set among all comparable single-model, Lidar-free methods. Our model achieves the highest Soft mAP of 47.32% and mAP of 46.24% on the test set; additionally, on the validation set, it also achieves the highest mAP of 46.37%. These results demonstrate the proposed model's ability to generate accurate and well-calibrated trajectory predictions. Notably, it also achieves the lowest displacement errors (0.5690 minADE and 1.1596 minFDE) and the best Overlap Rate (12.43%), indicating superior geometric accuracy and physical realism. Compared to its backbone, BeTopNet, HeLoFusion provides a clear improvement across most metrics, validating that our locality-focused approach effectively enhances feature representation. The strong performance confirms that modeling localized, heterogeneous interactions is a highly effective strategy for improving autonomous driving motion prediction accuracy. We also conducted a comparison of peak GPU memory usage to verify the computational scalability of our method, with more details available in Appendix B.

5 Conclusion

In this paper, we introduced HeLoFusion, a modular and efficient encoder for trajectory prediction in autonomous driving. Our approach is built on the principle of spatial locality, employing multi-scale local graphs to capture both pairwise and group-wise interactions, while explicitly handling agent heterogeneity, thereby producing powerful, socially aware agent representations. Our state-of-the-art results on the WOMD validate that the proposed design is a highly effective and practical strategy, offering an efficient module suitable for real-world autonomous driving systems.

Acknowledgments and Disclosure of Funding

This work was supported by National Key R&D Program of China (Grant No.2022ZD0160305).

References

Sergio Casas, Cole Gulino, Simon Suo, Katie Luo, Renjie Liao, and Raquel Urtasun. Implicit latent variable model for scene-consistent motion forecasting. In *European Conference on Computer Vision*, pages 624–641. Springer, 2020.

Alexander Cui, Sergio Casas, Kelvin Wong, Simon Suo, and Raquel Urtasun. Gorela: Go relative for viewpoint-invariant motion forecasting. *arXiv preprint arXiv:2211.02545*, 2022.

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.

Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9710–9719, 2021.

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.

Mohamed Hasan, Albert Solernou, Evangelos Paschalidis, He Wang, Gustav Markkula, and Richard Romano. Maneuver-aware pooling for vehicle trajectory prediction. *arXiv preprint arXiv:2104.14079*, 2021.

Xiaosong Jia, Penghao Wu, Li Chen, Yu Liu, Hongyang Li, and Junchi Yan. Hdgt: Heterogeneous driving graph transformer for multi-agent trajectory prediction via scene encoding. *IEEE transactions on pattern analysis and machine intelligence*, 45(11):13860–13875, 2023.

Chiyu Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, Dragomir Anguelov, et al. Motion-diffuser: Controllable multi-agent motion prediction using diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9644–9653, 2023.

Stepan Konev, Kirill Brodt, and Artsiom Sanakoyeu. Motioncnn: A strong baseline for motion prediction in autonomous driving. *arXiv preprint arXiv:2206.02163*, 2022.

Longzhong Lin, Xuewu Lin, Tianwei Lin, Lichao Huang, Rong Xiong, and Yue Wang. Eda: Evolving and distinct anchors for multimodal motion prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 3432–3440, 2024.

Haochen Liu, Li Chen, Yu Qiao, Chen Lv, and Hongyang Li. Reasoning multi-agent behavioral topology for interactive autonomous driving. *Advances in Neural Information Processing Systems*, 37:92605–92637, 2024.

Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. Multimodal motion prediction with stacked transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7577–7586, 2021.

Katie Luo, Sergio Casas, Renjie Liao, Xinchen Yan, Yuwen Xiong, Wenyuan Zeng, and Raquel Urtasun. Safety-oriented pedestrian occupancy forecasting. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1015–1022. IEEE, 2021.

Xiaoyu Mo, Zhiyu Huang, Yang Xing, and Chen Lv. Multi-agent trajectory prediction with heterogeneous edge-enhanced graph attention network. *IEEE Transactions on Intelligent Transportation Systems*, 23(7): 9554–9567, 2022.

Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S Refaat, and Benjamin Sapp. Wayformer: Motion forecasting via simple & efficient attention networks. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2980–2987. IEEE, 2023.

Jiquan Ngiam, Benjamin Caine, Vijay Vasudevan, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, et al. Scene transformer: A unified architecture for predicting multiple agent trajectories. *arXiv preprint arXiv:2106.08417*, 2021.

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*, pages 683–700. Springer, 2020.

Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S Refaat, Rami Al-Rfou, and Benjamin Sapp. Motionlm: Multi-agent motion forecasting as language modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8579–8590, 2023.

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.

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.

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*, 46(5):3955–3971, 2024.

Jiawei Sun, Jiahui Li, Tingchen Liu, Chengran Yuan, Shuo Sun, Zefan Huang, Anthony Wong, Keng Peng Tee, and Marcelo H Ang Jr. Rmp-yolo: A robust motion predictor for partially observable scenarios even if you only look once. *arXiv preprint arXiv:2409.11696*, 2024a.

Jiawei Sun, Chengran Yuan, Shuo Sun, Shanze Wang, Yuhang Han, Shuailei Ma, Zefan Huang, Anthony Wong, Keng Peng Tee, and Marcelo H Ang. Controlmtr: Control-guided motion transformer with scene-compliant intention points for feasible motion prediction. In *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1507–1514. IEEE, 2024b.

Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen, Bertrand Douillard, Chi Pang Lam, Dragomir Anguelov, et al. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7814–7821. IEEE, 2022.

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.

Qi Yan, Brian Zhang, Yutong Zhang, Daniel Yang, Joshua White, Di Chen, Jiachao Liu, Langechuan Liu, Binnan Zhuang, Shaoshuai Shi, et al. Trajflow: Multi-modal motion prediction via flow matching. *arXiv preprint arXiv:2506.08541*, 2025.

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.

Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17863–17873, 2023.

Table 2: **Comparison of Peak GPU Memory Usage During Training (Unit: MB)**. The "w/ Locality" column corresponds to the memory consumption of HeLoFusion (with spatial locality), and the "w/o Locality" column corresponds to that of the global method without spatial locality.

#Traffic Participants	w/ Locality	w/o Locality
29	2079.67	2166.51
92	2170.80	4918.78
122	2109.16	9984.98
165	2600.44	23223.32
235	2875.38	64858.23

A Implementation Details

Framework and Architecture. Our model is built upon the BeTopNet framework (Liu et al. [2024]), replacing its original encoder with our proposed HeLoFusion module while keeping the decoder architecture unchanged for a fair comparison. Following the protocol of MTR (Shi et al. [2022]), we use 64 predefined intention points as prediction anchors, which are generated via k-means clustering on the training set. Within the HeLoFusion encoder, the multi-scale interaction module constructs pairwise graphs from the $K = 10$ nearest neighbors and group-wise interactions using two hypergraphs with hyperedge sizes $S^{(1)}$ of 5 and 7. The subsequent context fusion module uses a local attention mechanism with a neighborhood size of 16. To account for agent heterogeneity, we employ three category-specific MLPs for the vehicle, pedestrian, and cyclist classes in the WOMD dataset.

Training. The model is trained for 30 epochs on 8 NVIDIA GeForce RTX 3090 GPUs. We use the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 0.01. The learning rate is managed by a step scheduler, decaying by a factor of 0.5 at epochs 22, 24, 26, and 28. We use a batch size of 80 and apply gradient accumulation to accommodate memory constraints. For training stability, gradient clipping is applied with a maximum norm of 1000.0. The network is initialized with weights from a BeTopNet model pretrained for 30 epochs.

B Computational Resource Requirement Analysis

To verify the computational scalability of HeLoFusion enabled by its spatial locality design, we compare the peak GPU memory usage during training between our method that leverages spatial locality and the global method that does not exploit spatial locality. All other experimental settings are kept identical between the two methods. In the experiment, we fix the number of target agents to predict at 4, systematically vary the total number of traffic participants in the scene, and measure the peak GPU memory usage of both methods during training.

As shown in Table 2, as the number of traffic participants in the scene increases, the peak memory of the global method without spatial locality grows exponentially, while the memory consumption of HeLoFusion exhibits a significantly more moderate growth trend. This result verifies that our spatial locality modeling achieves scalability in memory usage. Even in complex traffic scenarios with massive participants, HeLoFusion maintains low computational resource demands—a critical advantage for the practical deployment of autonomous driving systems.

C Limitations

Our work, while promising, has certain limitations. The proposed model has been validated on the Waymo Open Motion Dataset; its generalization to different datasets with unique traffic patterns and environmental conditions remains to be explored. And like other data-driven methods, its performance on rare, out-of-distribution events is not guaranteed.

D Broader Impacts Statement

The primary positive societal impact of our research is its potential to enhance autonomous driving safety, leading to fewer traffic accidents and improved transportation efficiency. However, we acknowledge potential negative consequences. The advancement of autonomous systems raises ethical questions regarding decision-making in unavoidable collisions. It is crucial that the development of such technology is paired with rigorous safety validation, research into algorithmic fairness to prevent biases, and thoughtful public policy to manage its societal transition.