MGFI: Multi-graph Fusion and Future Interaction-Aware Network for Heterogeneous Trajectory Prediction

Changxun Zhu[1] and Shibei Xue[1]

Abstract— **Trajectory prediction serves as a pivotal component within the autonomous driving technology stack. However, predicting trajectories in heterogeneous traffic environments remains a formidable challenge, primarily due to the intricate interactions among agents and the significant distinctness in kinematic patterns across different object categories. To address these issues, we propose a novel trajectory prediction framework named Multi-graph Fusion and Future Interaction-Aware Network (MGFI). Specifically, we first construct a multi-graph spatio-temporal feature encoding module designed to comprehensively capture complex interaction information among traffic participants. This module employs four independent graph neural network branches on heterogeneous nodes. To synthesize the node features derived from these branches, we introduce a hierarchical feature fusion module that selectively aggregates features from the aforementioned multi-graph features. Finally, the fused node representations are fed into a sequence prediction module, which is further augmented by anticipated future interaction features to enhance prediction accuracy. Extensive experiments conducted on ApolloScape dataset demonstrate that the proposed MGFI significantly outperforms state-of-the-art trajectory prediction algorithms in terms of both prediction accuracy and robustness.**

Keywords—Autonomous driving, Heterogeneous trajectory prediction, Deep learning.

I. INTRODUCTION

With the advent of the era of intelligence, a growing number of industries are undergoing a profound intelligent transformation [1]–[3]. As a strategic frontier in this era, autonomous driving technology has garnered significant attention due to its potential to reshape modern mobility patterns by mitigating human errors and substantially enhancing road safety and traffic efficiency [4]. To navigate safely and efficiently within complex real-world environments, autonomous vehicles (AVs) require not only precise perception capabilities but also the foresight to anticipate the future states of surrounding dynamic agents [5], [6]. In this context, trajectory prediction serves as a pivotal bridge connecting the upstream perception module and the downstream decision-making module [7].

Early research on trajectory prediction primarily focused on homogeneous environments, where agents in the scene possess similar dynamic properties and behavioral patterns, such as homogeneous crowds or highway traffic flows [8]–[11]. However, in highly heterogeneous urban traffic scenarios, the complexity of the task increases significantly. Urban traffic is not merely a collection of vehicles but a mixed space encompassing cars, buses, non-motorized vehicles, and pedestrians. These heterogeneous agents possess distinct kinematic properties and behavioral intentions. Additionally, semantic interaction challenges also exist, as the interaction logic among traffic participants in heterogeneous environments is inherently diverse [12].

Drawing on experience from homogeneous environments, Chandra et al. [13] combine CNNs and LSTMs to propose TraPHic, a heterogeneous trajectory prediction network designed to extract ego-trajectory and interaction features. However, this network exhibits limitations in interaction modeling. To address this issue, Li et al. [14] employ a spatio-temporal graph to model heterogeneous agents on the road and feed the agent features into a category-specific LSTM encoder-decoder structures for prediction. More advanced models, such as Trajectron++ [15], models the scene as a directed spatio-temporal graph and explicitly introduce dynamic constraints to ensure the physical feasibility of predicted trajectories. Nevertheless, such methods are limited by a single graph network structure, which is often insufficient for modeling complex road environments. To fully reflect road conditions using graph structures, Fang et al. [16] adopt a dual-graph framework, utilizing driving risk and environmental semantic information as priors for future trajectory prediction. However, the graphs constructed in this method merely rely on prior knowledge and the dual-graph feature fusion strategy is relatively simplistic. Zhang et al. [17] attempt to employ Graph Attention Network as the node feature aggregation module, coupled with a Transformer to progressively generate predicted trajectories stepwise. Although this method improved prediction accuracy, the introduction of extensive attention computations severely constrains inference and training speeds, hindering practical deployment.

To address the aforementioned challenges, this paper proposes a novel network named MGFI. To effectively cope with complex interactions in heterogeneous environments, we introduce a Multi-Graph Spatio-Temporal Feature Encoding Module. This module constructs four distinct graph topologies across different interaction dimensions: Driving Scene Graph, Category Graph, Similarity Graph, and Adaptive Graph. Specifically, the first three graphs employ prior knowledge to explicitly model interaction relationships, whereas the Adaptive Graph utilizes a data-driven approach to capture latent, non-prior interaction semantics. Subsequently, to rationally and effectively synthesize features from these independent graph branches, we propose a Hierarchical Graph Feature Fusion Module (HGF). Finally, to mitigate the oversight of evolving interactions in the

[1]Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, P.R. China (e-mail: shbxue@sjtu.edu.cn).

future horizon, we design an Attention-Enhanced Sequence Prediction Module (AESPM). In this module, we introduce a sparse self-attention mechanism to dynamically update spatial correlations and ensure the spatial consistency of the predicted trajectories during the decoding phase.

The reminder of this paper is structured as follows: Section II presents the problem formulation for heterogeneous traffic trajectory prediction. The proposed model is elaborated in Section III. Section IV reports the experimental results and analysis. Finally, the paper is concluded in Section V.

II. PROBLEM DESCRIPTION

Precise trajectory forecasting empowers autonomous vehicles to transition from passive reaction to proactive planning in response to their surroundings, which is particularly crucial for improving their forward-looking capabilities in complex environments. Consider a traffic scene comprising N agents across K categories. At time step t, the spatial coordinates of the i-th agent of category c is defined as $s^t_{c_i} = (x^t_{c_i}, y^t_{c_i})$. We define the past T_{obs} time steps as the observation horizon and the future T_{pred} time steps as the prediction horizon. The past state sequence of agent c_i is denoted as:

$$\mathbf{x}_{c_i} = \left[s^{t-T_{obs}+1}_{c_i}, s^{t-T_{obs}+2}_{c_i}, \ldots, s^t_{c_i}\right]. \tag{1}$$

The collection of past trajectories for all agents is represented as:

$$\mathbf{X} = \{\mathbf{x}_{c_i} \mid i \in \{1, \ldots, N_c\}, c \in \{1, \ldots, K\}\}. \tag{2}$$

Similarly, the set of future predicted trajectories is denoted as:

$$\mathbf{Y} = \{\mathbf{y}_{c_i} \mid i \in \{1, \ldots, N_c\}, c \in \{1, \ldots, K\}\}, \tag{3}$$

where $\mathbf{y}_{c_i} = \left[s^{t+1}_{c_i}, s^{t+2}_{c_i}, \ldots, s^{t+T_{pred}}_{c_i}\right]$ represents the predicted trajectory sequence of agent c_i .

In summary, this paper aims to predict future trajectories by leveraging the past trajectory information and corresponding category information of multi-agents. Given the inherent uncertainty in future movements, we employ a conditional probability model to formulate the predicted trajectories. The probability distribution of the future trajectories can be written as $P(\mathbf{Y} \mid \mathbf{X})$. In this paper, we set the number of agent categories to $K = 3$, focusing primarily on environments comprising vehicles, bicycles and pedestrians.

III. METHOD

As illustrated in Figure 1, the proposed MGFI network primarily consists of three modules. First, we construct multi-graph to extract complex spatial-temporal interaction features. Second, the multi-graph features are fused through hierarchical feature fusion module. Finally, we employ ConvGRU to construct the sequence prediction module. The remainder of this section details the overall network design and the specific implementation of each constituent module.

A. Multi-Graph Spatio-Temporal Feature Encoding Module

1) Construction of Multi-Graph: We model the past features of traffic participants as a unified spatio-temporal graph sequence $\mathscr{G} = \{\mathscr{G}_1, \mathscr{G}_2, \ldots, \mathscr{G}_T\}$, where T denotes the length of the historical sampling horizon, and $\mathscr{G}_t$ represents the spatial graph structure at time step t. Inspired by Ref. [18], we design four distinct graph topologies to comprehensively extract the latent interactions among traffic participants. In the following, we will describe the construction of the four proposed graph structures in detail. For simplicity, we omit the time step superscript.

First, we propose a Driving Scene Graph (DSG) to intuitively depict traffic scenarios. In the DSG, we incorporate the influence of distance and field of view among traffic participants on their future trajectories. By integrating these factors, the calculation formula for edge weights in the DSG can be formulated as follows:

$$a^S_{c_i c'_j} = \begin{cases} \dfrac{\mathbf{O}_{c_i c'_j}}{e^{\left\|\mathbf{P}_{c_i} - \mathbf{P}_{c'_j}\right\|/\sigma} - 1}, & if \left\|\mathbf{P}_{c_i} - \mathbf{P}_{c'_j}\right\| < d_D \\ 0, & otherwise. \end{cases} \tag{4}$$

where $\mathbf{P}_{c_i} = \left(x_{c_i}, y_{c'_i}\right)$ denotes the position of agent c_i and $\left\|\mathbf{P}_{c_i} - \mathbf{P}_{c'_j}\right\|$ denotes the Euclidean distance between agent c_i and agent c'_j. The decay rate hyperparameter σ is constrained such that $0 < \sigma < \frac{d_D}{2}$. $\mathbf{O}_{c_i c'_j}$ represents the visibility coefficient of agent c'_j relative to agent c_i. The value of $\mathbf{O}_{c_i c'_j}$ is set to $\cos\phi_i$ if $\phi_i \in [-\pi/2, \pi/2]$ and 0 otherwise, where ϕ_{c_i} is the angle between the relative position vector $\overrightarrow{c_i c'_j}$ and agent c_i's heading. The field of view is modeled as a 180° semicircle symmetric about the motion vector.

To capture the categorical characteristics inherent in heterogeneous traffic environments, we introduce a Category Graph (CG). The edge weights of the category graph are defined by the following formulation:

$$a^C_{c_i c'_j} = \begin{cases} 1, & if\ c = c' \\ 0, & otherwise. \end{cases} \tag{5}$$

Through the structural constraint, this graph emphasizes the similarity in driving intentions among agents with the same class.

While the CG effectively captures intention similarity within the same class, heterogeneous traffic settings often exhibit shared motion intentions across different classes. In this case, we introduce a Similarity Graph (SG) to capture shared motion patterns among diverse agents. By leveraging the similarity of historical trajectories, the edge weights calculated as follows:

$$a^{Sim}_{c_i c'_j} = \begin{cases} \dfrac{\mathbf{Sim}\left(\mathbf{x}_{c_i}, \mathbf{x}_{c'_j}\right)}{\left\|\mathbf{P}_{c_i} - \mathbf{P}_{c'_j}\right\|} & if \left\|\mathbf{P}_{c_i} - \mathbf{P}_{c'_j}\right\| < d_D \\ 0, & otherwise. \end{cases} \tag{6}$$

where $\mathbf{x}_{(\cdot)} \in \mathbb{R}^{C \times T}$ represents the past trajectory sequence vector of a traffic agent, $\mathbf{Sim}(\cdot, \cdot)$ measures the similarity between the past trajectory sequences of different agents

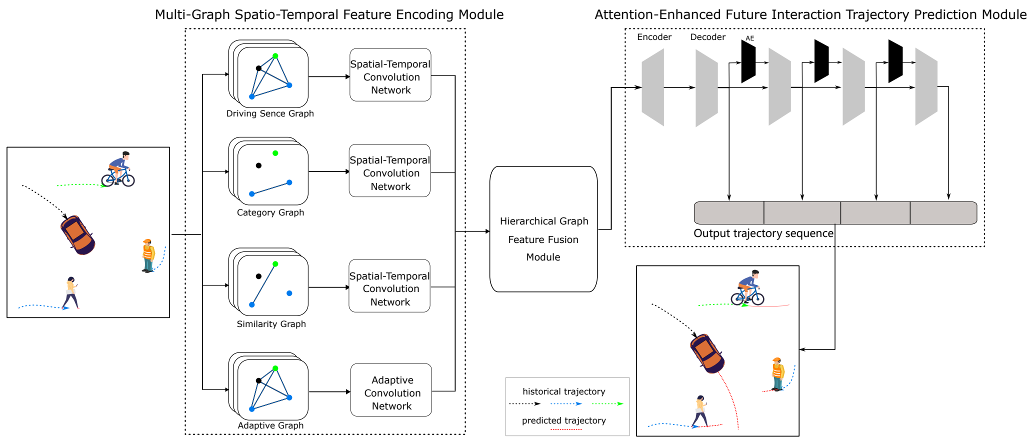

Fig. 1 Schematic diagram of heterogeneous trajectory prediction network structure.

and we employ cosine similarity for this computation. Following the construction of the prior graphs, separate Spatial-Temporal Graph Convolutional Networks are deployed within each branch for node feature aggregation.

With three prior-based graphs mentioned above, complex heterogeneous interactions may still cause certain key features to be inadequately modeled or fall outside the scope of the pre-defined priors. To extract these implicit features and compensate for the structural limitations of the prior-based graph networks, inspired by [19], we introduce an adaptive graph as a supplement to the structured graphs constructed from prior knowledge. The adjacency matrix of the adaptive graph is constructed using two learnable node embedding matrices $\mathbf{E}_1$ and $\mathbf{E}_1$, formulated as follows:

$$\widetilde{A}^{adp} = D^{-\frac{1}{2}} A^{adp} D^{-\frac{1}{2}} = \mathrm{softmax}(ReLu(\mathbf{E}_1 \mathbf{E}_2{}^{T})). \tag{7}$$

2) Hierarchical Graph Feature Fusion Module: To effectively exploit multiple prior structures while mitigating noise from the learned adaptive graph, we propose a hierarchical graph feature fusion module. Let $\{\mathbf{F}^k\}_{k\in\{S,C,Sim\}}$ and $\mathbf{F}^{adp}$ denote the node features derived from the prior graph branches and the adaptive branch, respectively. All features share the shape $\mathbb{R}^{N\times C\times T\times V}$. We first aggregate the prior branches via an adaptive weighting mechanism to secure a stable representation, followed by a residual correction using the adaptive features. The detailed operations of this module are presented as follow.

First, we project the heterogeneous features from each prior branch into a unified subspace via a 1×1 convolution. To capture dynamic spatiotemporal dependencies, we compute position-wise fusion weights α based on the concatenated representations:

$$\tilde{\mathbf{X}}^k = \mathbf{W}_k * \mathrm{Norm}(\mathbf{F}^k), \quad k \in \{S, C, Sim\}, \tag{8}$$

$$\alpha = \mathrm{softmax}(\mathbf{W}_2 * \delta\left(\mathbf{W}_1 * \left[\tilde{\mathbf{X}}^S \parallel \tilde{\mathbf{X}}^C \parallel \tilde{\mathbf{X}}^{Sim}\right]\right)), \tag{9}$$

$$\mathbf{F}_{prior} = \sum_k \alpha_k \odot \tilde{\mathbf{X}}^k, \tag{10}$$

where $\parallel$ denotes concatenation, δ is the ReLU activation, and $\mathbf{W}_{(\cdot)}$ are learnable parameters. This mechanism ensures the model selectively focuses on the most informative prior interaction types at each spatiotemporal location.

While prior graphs provide structural stability, they may miss latent interactions in complex scenarios. To address this, we incorporate the adaptive graph feature $\mathbf{F}^{adp}$ as a residual supplement. We design a completion gate $\gamma \in [0,1]$ to regulate the magnitude of this supplementation:

$$\tilde{\mathbf{X}}^{adp} = \mathbf{W}_{adp} * \mathrm{Norm}(\mathbf{F}^{adp}), \tag{11}$$

$$\gamma = \sigma\left(f_\gamma\left(\left[\mathbf{F}_{prior} \parallel \tilde{\mathbf{X}}^{adp}\right]\right)\right), \tag{12}$$

$$\mathbf{F}_{fusion} = \mathbf{F}_{prior} + \gamma \odot \tilde{\mathbf{X}}^{adp}, \tag{13}$$

where σ is the sigmoid function and f_γ follows the same structure as the gating network in Eq. (9). This gated residual connection allows the adaptive branch to specifically compensate for deficiencies in the prior knowledge without introducing noise during early training stages.

B. Attention-Enhanced Sequence Prediction Module

1) ConvGRU-Based Trajectory Prediction Module: To effectively capture complex temporal dependencies while preserving the spatial topology of fused features, we employ a ConvGRU-based Sequence-to-Sequence (Seq2Seq) architecture as the trajectory prediction module at the end of the network. Specifically, the module comprises an Encoder and a Decoder, both of which are composed of stacked ConvGRU layers. The Encoder map the historical input sequence into hidden state tensors containing deep spatiotemporal semantics. The Decoder uses these tensors as initial states to progressively infer future prediction sequences via recursive generation. Following Ref. [18], we adopt independent prediction modules for different categories of traffic participants to mitigate the impact of class heterogeneity on prediction accuracy. The feature calculation for a specific category of traffic participants is formulated as follows:

$$\mathscr{H}^{enc} = \text{Encoder}\left(\mathbf{F}_{fusion}\right), \tag{14}$$

$$\mathscr{H}_t^{dec} = \text{Decoder}\left(\hat{\mathbf{Y}}_{t-1}, \mathscr{H}_{t-1}^{dec}\right), \tag{15}$$

where $\hat{\mathbf{Y}}_{t-1}$ denotes the embedding of the predicted position from the previous step. The final decoded tensor $\mathbf{F}^{dec} \in \mathbb{R}^{N\times C^{dec}\times T_{pred}}$ is constructed by stacking the decoded features across all prediction time steps. Finally, we apply a fully connected layer to generate the parameters of the trajectory probability distribution.

2) Attention-Enhanced Module: As indicated in the description of the prediction module, the encoder is responsible for extracting temporal features, while the decoder generates predicted trajectories step-by-step based on these features. However, this configuration neglects the interactions among traffic agents at the already generated prediction timesteps.

To address this limitation, we propose an Attention-Enhanced Module (AE) to capture interaction features among agents at the previous prediction timestep and integrate them into the decoder's trajectory generation via a gating mechanism, thereby enhancing spatial dependencies during temporal prediction. Unlike standard dense attention, we introduce a sparse attention mechanism to model the inter-agent interactions at the last prediction time step. Let $\mathbf{F}_{t-1}^{dec} \in \mathbb{R}^{N\times C}$ denote the decoder output from the previous time step, representing the latent states of N agents. We first project $\mathbf{F}_{t-1}^{dec}$ into queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$. The raw attention matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ is computed via the scaled dot-product:

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right). \tag{16}$$

To eliminate redundant interactions, we propose a learnable masking strategy. The raw attention matrix $\mathbf{A}$ is first compressed via a 1×1 convolution layer followed by a ReLU activation to obtain a feature map $\mathbf{F}_A$. Subsequently, we generate a binary sparsity mask $\mathbf{M}$ based on a predefined threshold ξ:

$$\mathbf{M}_{ij} = \mathbb{I}\left(\sigma(\mathbf{F}_A)_{ij} > \xi\right), \tag{17}$$

where $\sigma(\cdot)$ is the sigmoid function and $\mathbb{I}(\cdot)$ is the indicator function. This mask retains only significant interactions. The sparsified attention output $\tilde{\mathbf{X}}_{sp}$ is then computed by applying the mask to the attention weights and aggregating the values:

$$\tilde{\mathbf{X}}_{sp} = \text{LayerNorm}\left((\mathbf{M}\odot\mathbf{A})\mathbf{V} + \mathbf{F}_{t-1}^{dec}\right). \tag{18}$$

Finally, the enhanced feature $\tilde{\mathbf{X}}_{sp}$ is concatenated with the input of the next prediction step, enriching the trajectory generation with sparse yet salient social context.

IV. EXPERIMENTS AND RESULTS

A. Datasets

To validate the accuracy and reliability of our proposed model for heterogeneous traffic trajectory prediction, we utilize ApolloScape [20] dataset for experiment. Following the data processing protocol in existing work [21], we define a 3 seconds historical observation window to predict the future trajectories of each agent over the next 3 seconds.

B. Evaluation Metrics

Average Displacement Error (ADE) and Final Displacement Error (FDE) are common metrics for homogeneous trajectory prediction tasks. Building upon these, we further employ the Weighted Sum of ADE (WSADE) and Weighted Sum of FDE (WSFDE) to accurately evaluate the model's prediction performance for diverse traffic participants in heterogeneous environments. Following the weight assignment in [20], the error weights for vehicles, bicycles, and pedestrians are set to 0.20, 0.22, and 0.58, respectively. Consequently, the formulations for WSADE and WSFDE are expressed as:

$$WSADE = 0.2\times ADE_{veh} + 0.58\times ADE_{ped} + 0.22\times ADE_{bike}, \tag{19}$$

$$WSFDE = 0.2\times FDE_{veh} + 0.58\times FDE_{ped} + 0.22\times FDE_{bike}. \tag{20}$$

C. Implementation Details

Environment setting. The proposed heterogeneous traffic trajectory prediction model is developed using the PyTorch framework (version 2.4.0). In the training phase, the batch size B is set to 64, and the total number of training epochs R is set to 150. We utilize the Adam optimizer to update model parameters with an initial learning rate of $LR = 0.01$. To accelerate the convergence of the loss function, we adopt a multi-step learning rate decay strategy. The learning rate is decayed every 40 epochs with a decay factor of $\gamma = 0.5$. The entire training process is conducted on an NVIDIA GTX 3060 GPU.

Loss function. Considering the inherent uncertainty of future trajectories, we formulate the predicted coordinates of each traffic participant at any given future time step as random variables governed by a bivariate Gaussian distribution. Under this probabilistic framework, the model is optimized by minimizing the negative log-likelihood (NLL) of the ground-truth observations. Formally, the loss function is defined as:

$$Loss = -\sum_{n=1}^{N}\sum_{t=T+1}^{T+t_f}\left(\mathbb{P}\left(x_n^t, y_n^t \mid \widehat{\mu}_n^t, \widehat{\sigma}_n^t, \widehat{\rho}_n^t\right)\right). \tag{21}$$

D. Experiments

1) Comparative Analysis: To showcase the superiority of the proposed framework in heterogeneous trajectory prediction, we benchmark it against the following state-of-the-art methods in this subsection.

- Social LSTM [8]: This baseline model accounts for inter-agent interactions by introducing a social pooling layer.
- CS-LSTM [22]: This baseline model adopts an overall "encoder-decoder" architecture, integrating LSTM, CNN, and social interactions.
- GRIP++ [14]: This baseline employs a fixed graph convolutional blocks to extract spatio-temporal features.
- AI-TP [23]: This baseline model employs a graph attention model to construct agent interactions. It weights

TABLE I: Comparison with baselines on the ApolloScape dataset. Data are in meters.

Model	ADE_v	ADE_p	ADE_b	WSADE	FDE_v	FDE_p	FDE_b	WSFDE
Social LSTM	2.95	1.29	2.56	1.89	5.28	2.32	4.54	3.40
CS-LSTM	-	-	-	2.14	-	-	-	4.70
GRIP++	2.24	0.71	1.80	1.25	4.07	1.37	3.41	2.36
AI-TP	1.98	0.67	1.71	1.15	3.52	1.26	3.14	2.13
D2-TPred	-	-	-	1.02	-	-	-	1.69
ours	**1.58**	**0.62**	**1.29**	**0.95**	**2.65**	**1.01**	**2.09**	**1.56**

neighbor nodes in a data-driven manner to highlight critical interactions.

- D2-TPred [24]: This baseline model utilizes a heterogeneous graph neural network to enhance the interaction between traffic participants and the environment.

The prediction accuracy of the proposed model and the baseline models on the ApolloScape dataset is reported in Table I. As shown in the Table I, our proposed model outperforms almost all existing models. Compared to D2-TPred, the second-best performing model, our model still shows considerable improvements in both weighted accuracy metrics. Observing the metrics across different categories, the performance gain for vehicles and bicycles is notably higher than that for pedestrians. This is because bicycles and vehicles move at relatively higher speed, making their patterns more distinct in the data. Consequently, the weighted average error derived from different categories effectively reflects the overall model performance. Due to space limitations, detailed experimental results and analysis on the nuScenes dataset are provided in the extended arXiv version of this paper.

2) Ablation Study: To further demonstrate the effectiveness of the individual modules within the network proposed in this paper, we conduct ablation studies in this section. These experiments aim to validate the positive contribution of the relevant modules to the model's prediction performance. We trained distinct model variants by removing specific modules or their key components. The detailed configurations of each variants are listed in Table II.

TABLE II: Comparison of the WSADE metrics for the ablation experiments on the ApolloScape dataset. Data are in meters.

Index	DSG	CG	SG	AG	HGF	SPM	AE	WSADE
A1	√	×	×	×	×	×	×	1.36
A2	√	√	×	×	×	×	×	1.25
A3	√	√	√	×	×	×	×	1.17
A4	√	√	√	√	×	×	×	1.12
A5	√	√	√	√	√	×	×	1.04
A6	√	√	√	√	√	√	×	0.98
A7	√	√	√	√	√	√	√	0.95

As demonstrated by the results of model variants A1 to A7, the progressive integration of each module leads to a consistent reduction in prediction error. Results from variants A1 to A3 highlight the role of prior graph branches in lowering prediction errors. By introducing an adaptive graph, variant A4 captures the heterogeneous interactions omitted by prior models, yielding further error reductions. Variant A5 refines the feature fusion mechanism by adopting a hierarchical fusion strategy to integrate multi-graph features, ultimately generating the fused node representations. Lastly, the attention-enhanced sequence prediction module further improves the model's prediction performance.

3) Qualitative Analysis: This subsection visually analyzes the model's prediction performance in typical interaction scenarios to intuitively demonstrate its trajectory prediction capability when handling multi-agent interactions. Given that the model outputs a probability distribution of trajectories, the mean of the distribution is adopted as the final predicted trajectory to facilitate a direct comparison with the ground truth and the baseline model (GRIP++).

The visualization results are presented in Fig. 2, covering six representative time frames. Overall, the proposed model demonstrates robust prediction performance, outperforming the GRIP++ model in prediction accuracy for various traffic participants across most scenarios. Fig. 2(a)-(b) depict heterogeneous interaction scenarios involving mixed interactions among vehicles, pedestrians, and cyclists. Fig. 2(c) illustrates a scenario where pedestrians cross a two-way road. The model accurately captures the pedestrians' crossing intentions. Fig. 2(d)-(e) depict car-following and parallel driving scenarios within a homogeneous vehicle environment. The results indicate that the model accurately predicts the future trajectories of these vehicles, demonstrating strong kinematic consistency. Fig. 2(f) features a collision avoidance interaction. In summary, the visualization results fully validate the accuracy and rationality of the proposed model in predicting trajectories for diverse traffic participants within complex and dynamic traffic environments.

V. CONCLUSIONS

In this paper, we have proposed a novel framework named MGFI to address the challenges of trajectory prediction in heterogeneous traffic environments. By constructing a multi-graph spatio-temporal feature encoding module, we explicitly model the complex interactions among diverse agents through four specific graphs. To effectively synthesize these features, we propose a hierarchical fusion module to combine prior knowledge with data-driven adaptive features. Furthermore, we design an attention-enhanced sequence prediction module to capture dynamic spatial dependencies during the future trajectory generation process. Extensive experiments on the ApolloScape dataset demonstrate that our MGFI significantly outperforms state-of-the-art methods in terms of prediction accuracy and robustness.

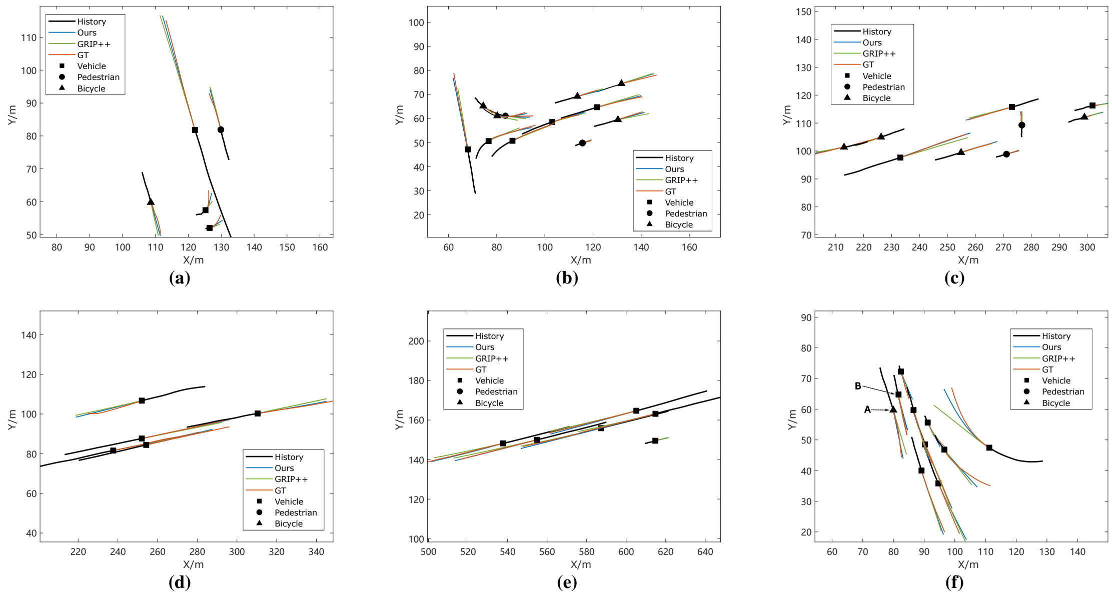

Fig. 2 Schematic of visualized prediction results

REFERENCES

[1] Q. Guan, W. Li, S. Xue, and D. Li, "High-resolution representation object pose estimation from monocular images," in *2021 China Automation Congress (CAC)*. IEEE, 2021, pp. 980–984.

[2] Q. Guan, Z. Sheng, and S. Xue, "Hrpose: Real-time high-resolution 6d pose estimation network using knowledge distillation," *Chinese Journal of Electronics*, vol. 32, no. 1, pp. 189–198, 2023.

[3] Q. Wu, L. Liu, and S. Xue, "Global update guided federated learning," in *2022 41st Chinese Control Conference (CCC)*. IEEE, 2022, pp. 2434–2439.

[4] Y. Gao, K. Yang, Y. Yue, and Y. Wu, "A vehicle trajectory prediction model that integrates spatial interaction and multiscale temporal features," *Scientific Reports*, vol. 15, no. 1, p. 8217, 2025.

[5] H. Li, Z. Liao, Y. Rui, L. Li, and B. Ran, "A physical law constrained deep learning model for vehicle trajectory prediction," *IEEE Internet of Things Journal*, vol. 10, no. 24, pp. 22 775–22 790, 2023.

[6] Z. Sheng, L. Liu, S. Xue, D. Zhao, M. Jiang, and D. Li, "A cooperation-aware lane change method for automated vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 3, pp. 3236–3251, 2022.

[7] Y. Huang, J. Du, Z. Yang, Z. Zhou, L. Zhang, and H. Chen, "A survey on trajectory-prediction methods for autonomous driving," *IEEE transactions on intelligent vehicles*, vol. 7, no. 3, pp. 652–674, 2022.

[8] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961–971.

[9] X. Li, X. Ying, and M. C. Chuah, "Grip: Graph-based interaction-aware trajectory prediction," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 3960–3966.

[10] Z. Sheng, Y. Xu, S. Xue, and D. Li, "Graph-based spatial-temporal convolutional network for vehicle trajectory prediction in autonomous driving," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 10, pp. 17 654–17 665, 2022.

[11] X. Mo, Y. Xing, and C. Lv, "Graph and recurrent neural network-based vehicle trajectory prediction for highway driving," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 1934–1939.

[12] Y. Liu, B. Li, X. Wang, C. Sammut, and L. Yao, "Attention-aware social graph transformer networks for stochastic trajectory prediction," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 11, pp. 5633–5646, 2024.

[13] R. Chandra, U. Bhattacharya, A. Bera, and D. Manocha, "Traphic: Trajectory prediction in dense and heterogeneous traffic using weighted interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8483–8492.

[14] X. Li, X. Ying, and M. C. Chuah, "Grip++: Enhanced graph-based interaction-aware trajectory prediction for autonomous driving," *arXiv preprint arXiv:1907.07792*, 2019.

[15] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *European Conference on Computer Vision*. Springer, 2020, pp. 683–700.

[16] J. Fang, C. Zhu, P. Zhang, H. Yu, and J. Xue, "Heterogeneous trajectory forecasting via risk and scene graph learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 11, pp. 12 078–12 091, 2023.

[17] K. Zhang, X. Feng, L. Wu, and Z. He, "Trajectory prediction for autonomous driving using spatial-temporal graph attention transformer," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 22 343–22 353, 2022.

[18] Z. Sheng, Z. Huang, and S. Chen, "Ego-planning-guided multi-graph convolutional network for heterogeneous agent trajectory prediction," *Computer-Aided Civil and Infrastructure Engineering*, vol. 39, no. 22, pp. 3357–3374, 2024.

[19] L. Bai, L. Yao, C. Li, X. Wang, and C. Wang, "Adaptive graph convolutional recurrent network for traffic forecasting," *Advances in neural information processing systems*, vol. 33, pp. 17 804–17 815, 2020.

[20] Y. Ma, X. Zhu, S. Zhang, R. Yang, W. Wang, and D. Manocha, "Trafficpredict: Trajectory prediction for heterogeneous traffic-agents," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 6120–6127.

[21] L. Fang, Q. Jiang, J. Shi, and B. Zhou, "Tpnet: Trajectory proposal network for motion prediction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6797–6806.

[22] N. Deo and M. M. Trivedi, "Convolutional social pooling for vehicle trajectory prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 1468–1476.

[23] K. Zhang, L. Zhao, C. Dong, L. Wu, and L. Zheng, "Ai-tp: Attention-based interaction-aware trajectory prediction for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 73–83, 2022.

[24] Y. Zhang, W. Wang, W. Guo, P. Lv, M. Xu, W. Chen, and D. Manocha, "D2-tpred: Discontinuous dependency for trajectory prediction under traffic lights," in *European Conference on Computer Vision*. Springer, 2022, pp. 522–539.