Towards explainable motion prediction using heterogeneous graph representations

Transportation Research Part C 157 (2023) 104405

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Towards explainable motion prediction using heterogeneous graph representations

Sandra Carrasco Limeros [a,c], Sylwia Majchrowska [b,c], Joakim Johnander [c,d], Christoffer Petersson [c,e], David Fernández Llorca [f,a,*]

[a] Computer Engineering Department, Polytechnic School, University of Alcala, Madrid, Spain
[b] AI Sweden, Göteborg, Sweden
[c] Zenseact AB, Göteborg, Sweden
[d] Department of Electrical Engineering, Linköping University, Linköping, Sweden
[e] Chalmers University of Technology, Göteborg, Sweden
[f] European Commission, Joint Research Centre, Seville, Spain

ARTICLE INFO

Keywords:
Autonomous vehicles
Explainable artificial intelligence
Heterogeneous graph neural networks
Multi-modal motion prediction

ABSTRACT

Motion prediction systems play a crucial role in enabling autonomous vehicles to navigate safely and efficiently in complex traffic scenarios. Graph Neural Network (GNN)-based approaches have emerged as a promising solution for capturing interactions among dynamic agents and static objects. However, they often lack transparency, interpretability and explainability — qualities that are essential for building trust in autonomous driving systems. In this work, we address this challenge by presenting a comprehensive approach to enhance the explainability of graph-based motion prediction systems. We introduce the Explainable Heterogeneous Graph-based Policy (XHGP) model based on an heterogeneous graph representation of the traffic scene and lane-graph traversals. Distinct from other graph-based models, XHGP leverages object-level and type-level attention mechanisms to learn interaction behaviors, providing information about the importance of agents and interactions in the scene. In addition, capitalizing on XHGP's architecture, we investigate the explanations provided by the GNNExplainer and apply counterfactual reasoning to analyze the sensitivity of the model to modifications of the input data. This includes masking scene elements, altering trajectories, and adding or removing dynamic agents. Our proposal advances towards achieving reliable and explainable motion prediction systems, addressing the concerns of users, developers and regulatory agencies alike. The insights gained from our explainability analysis contribute to a better understanding of the relationships between dynamic and static elements in traffic scenarios, facilitating the interpretation of the results, as well as the correction of possible errors in motion prediction models, and thus contributing to the development of trustworthy motion prediction systems.

The code to reproduce this work is publicly available at https://github.com/sancarlim/Explainable-MP/tree/v1.1.

1. Introduction

Autonomous vehicles (AVs) must perform trajectory planning based on the global route and the local context. Accurate and safe trajectory planning depends on the system's ability to anticipate the future motions of surrounding agents, a key challenge in achieving full self-driving autonomy (Bahari et al., 2021).

* Corresponding author at: European Commission, Joint Research Centre, Seville, Spain.
E-mail addresses: sandra.carrascol@uah.es (S. Carrasco Limeros), sylwia.majchrowska@ai.se (S. Majchrowska), joakim.johnander@zenseact.com (J. Johnander), christoffer.petersson@zenseact.com (C. Petersson), david.fernandez-llorca@ec.europa.eu (D. Fernández Llorca).

https://doi.org/10.1016/j.trc.2023.104405
Received 21 November 2022; Received in revised form 27 July 2023; Accepted 31 October 2023
Available online 6 November 2023

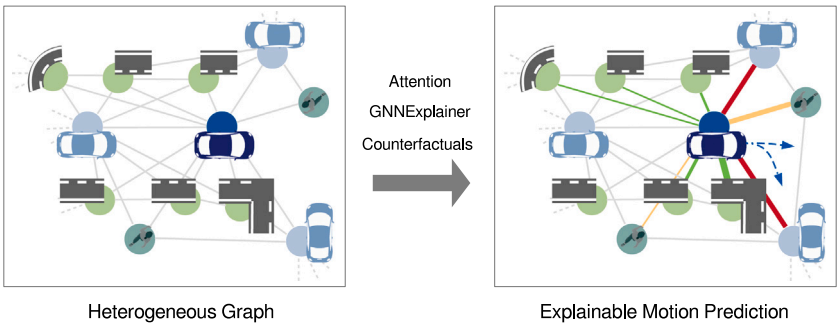

Fig. 1. Explainable Heterogeneous Graph-based Policy (XHGP). Explainable motion prediction is approached from three points of view: by visualizing the attention learned by the model, by using the GNNExplainer method, and by exploring counterfactual reasoning.

Graph-based approaches are gaining traction, as traffic scenarios can be naturally represented as graphs. Recent studies have extended deep learning techniques to graph data, such as Graph Convolutional Networks (GCNs) (Kipf and Welling, 2017) and Graph Attention Networks (GATs) (Veličković et al., 2018). These approaches incorporate information from node characteristics in a graph, such as position or velocity, as well as information on road structure. By transmitting messages across graph edges, these models can address complex and highly interactive scenarios with diverse road topologies and multiple types of agents (Carrasco et al., 2021; Deo et al., 2021; Gu et al., 2021).

Multi-modal motion prediction captures the underlying distribution of future motions, handling multiple predictions with associated probabilities. This approach produces more balanced and robust performance, improving interpretability when accompanied by intent detection systems and appropriate visualization (Carrasco Limeros et al., 2023).

Despite the successes of graph-based models, a significant challenge is their inherent complexity, making them opaque and often difficult to interpret. Regulators are increasingly demanding transparency requirements for artificial intelligence (EC, 2021; Office of U.S. Senator Ron Wyden, 2022; European Commission and Directorate-General for Communications Networks, Content and Technology, 2019), including traceability, data logging and explainability. In the context of autonomous driving, these requirements can benefit various stakeholders, such as internal and external users, testing and certification bodies and developers (Gonzalo et al., 2022; Fernández Llorca and Gómez, 2023). Recognizing this challenge, our work emphasizes not just accurate motion prediction but more crucially, the explainability of these predictions.

In this paper, we address the challenge of explainability in motion prediction systems by proposing an Explainable Heterogeneous Graph-based Policy (XHGP) model that is designed to improve the explainability of the motion prediction task (see Fig. 1). Our contributions are three-fold:

- We introduce the Explainable Heterogeneous Graph-based Policy (XHGP) model, an approach that distinctively combines the graph-traversals method from Deo et al. (2021) with the strengths of heterogeneous graph representation and attention (Yang et al., 2021), specifically architected to enhance explainability in motion prediction. Our method jointly models road and dynamic agents in a heterogeneous graph, offering increased transparency in motion prediction. We evaluate the performance of our model on the challenging large-scale nuScenes dataset, achieving competitive results compared to prior works.
- We analyze various explainability methods applied to Graph Neural Networks (GNNs) in the context of multi-modal motion prediction, examining two types of attention mechanisms, GNNExplainer, and counterfactual reasoning, providing a comprehensive assessment of explainability in graph-based motion prediction models.
- We conducted a qualitative and quantitative evaluation of the explanations provided using the diverse and challenging nuScenes dataset, demonstrating the effectiveness and potential of our approach in improving the transparency of motion prediction systems in autonomous driving.

By incorporating explainability into motion prediction systems, our work takes a significant step towards more transparent and reliable autonomous vehicles. This is crucial for user confidence, developer understanding, and regulatory compliance (Fernández Llorca and Gómez Gutierrez, 2021).

The remaining of this paper is structured as follows: Section 2 provides a comprehensive overview of related works, covering GNN-based motion prediction, explainability with Graph Neural Networks, and interpretable and explainable motion prediction. In Section 3, we introduce our methodology, starting with the problem formulation, then detailing our multi-modal motion prediction model, and finally elaborate on the applied explainability techniques. Explainable motion prediction is the core aspect of our work — we explore attention visualization, post-hoc explainability methods using GNNExplainer, and counterfactual reasoning. Section 4 presents the evaluation framework, including the dataset, and evaluation metrics used for our experiments. In Section 5, we discuss our results. Finally, Section 6 provides a discussion and conclusion of our work, summarizing our findings and outlining potential future research directions.

2. Related works

Deep neural networks have been successful in a wide range of tasks, including motion prediction (Huang et al., 2022). However, a well-known drawback with deep neural networks is that it is difficult to explain their predictions. Without reasoning about the underlying mechanisms behind the predictions, deep models cannot be fully trusted, which precludes their use in critical applications due to fairness, privacy, and safety concerns. The Defence Advanced Research Project Agency (DARPA) coined the term Explainable AI (XAI) as a new research direction to address this weakness (Gunning, 2016).

The terms of interpretability and explainability are usually interchangeably used. However, although closely related, some works try to clarify the difference between these concepts, Doshi-Velez and Kim (2017), Rudin (2019), Gilpin et al. (2018), but none of them provide a rigorous formal mathematical definition. *Interpretability* is defined by Doshi-Velez and Kim (2017) as "the ability to explain or to present in understandable terms to a human" meaning that the cause and effect can be determined. In this work, we follow (Rudin, 2019) and consider a model to be interpretable if the model itself can provide understandable interpretations of its predictions, such as a decision tree model. However, most machine learning models are designed with specific accuracy as a goal and do not have interpretability constraints. In contrast, *explainable* models are those that are still a black box and require post hoc techniques to explain its predictions. According to Doshi-Velez et al. (2017), an explanation must provide at least one of the following: the main factors in the decision process; whether the decision would have changed if the factors had been modified; whether two similar cases result in different decisions, or vice versa.

In this section, we first review recent advances in graph-based motion prediction. Next, we review explainability in GNNs. Finally, we identify previous attempts to deal with interpretable and explainable motion prediction.

2.1. GNN-based motion prediction

Graph Neural Networks (GNNs) have been widely used to capture and model the underlying spatial and temporal relationships of the agents in a traffic scene. Some approaches propose the use of the GNN model to encode the spatial interactions between traffic agents and then attaching some recurrent (Li et al., 2019; Zhou et al., 2021; Mo et al., 2021; Tang et al., 2023; Zhang et al., 2023) or attention-based (Zhang et al., 2022a) system to model the temporal relationships and generate the predicted trajectories. However, the temporal information can be intrinsically captured by the graph model by adding self-connections to the nodes as temporal edges (Carrasco et al., 2021).

Apart from the spatial and temporal information of the agents, prior knowledge of the road structure plays a fundamental role. Most self-driving cars have access to high definition (HD) vector maps, which contain detailed geometric information, such as roads, lanes, intersections, crossings, traffic signs, and traffic lights. The simplistic way to encode road and motion information is rasterization (Casas et al., 2020). In such a representation, different semantics are encoded in separate channels, which facilitates the learning of the convolutional neural network (CNN). However, rasterization may lose useful information like lane topologies and has difficulty to capture long range interactions. A more efficient solution is to represent structured HD maps and agents using polylines. Gao et al. (2020) treat a map as a collection of polylines and use self attention-subgraphs to encode them. Discretization of the scene context in form of vectors was adopted in subsequent works (Zhao et al., 2020; Gu et al., 2021; Zhao et al., 2022) because it provides a uniform representation with a lower computational cost. In addition, it allows interpretable analysis of the model's behavior and enables counterfactual predictions that condition hypothetical "what-if" polylines (Khandelwal et al., 2020).

Other ideas involve building additional structures. For example, LaneGCN (Liang et al., 2020) particularly emphasizes a graph of lanes and conducts convolutions over the graph. LaneGCN's architecture is deep and complex with multiple fusion modules and spatial attention layers, each targeting specific interactions. LaneRCNN (Zeng et al., 2021) encodes both actors and maps in an unified graph representation, which is even more structured. Both methods encode the input context into a single context vector. It is then used by the multi-modal prediction header to derive multiple likely future trajectories. The predictive header therefore has to learn a complex mapping, which often leads to predictions that get out of the way or violate traffic rules.

Recently introduced Prediction via Graph-based Policy (PGP) model (Deo et al., 2021) achieves great scene compliance due to the use of a lane-graph traversals approach. It selectively aggregates scene context based on path traversals sampled from a learned behavior cloning policy, capturing the lateral variability of the output distribution. To model longitudinal variability, a sampled latent variable was added, enabling prediction of different trajectories for a same traversal.

In contrast to the aforementioned models, our proposed XHGP adopts a heterogeneous graph representation, capturing a broader and richer context of the scene. This not only allows for comprehensive motion predictions but also facilitates enhanced explainability, as we will discuss in subsequent sections.

2.2. Explainability with graph neural networks

Explainability in GNNs can be taxonomized in instance-level and model-level methods. The former provide input-dependent explanations, while the latter explain the general behavior of a GNN at a higher level. XGNN (Yuan et al., 2020) proposes to explain GNNs by learning to generate graphs that achieve optimal prediction scores according to the GNN model to be explained. Instance-level methods have been studied much more extensively and can be divided into four categories:

- **Gradient-based methods** use gradient or hidden features as proxies for input importance to explain the predictions through back-propagation. (Pope et al., 2019; Baldassarre and Azizpour, 2019)

- **Perturbation-based methods** study the change of output variations for different input perturbations. (Ying et al., 2019; Luo et al., 2020a; Schlichtkrull et al., 2021; Lin et al., 2021; Lucic et al., 2021)
- **Decomposition-based methods** track the contribution of individual graph components to the final prediction by decomposing the original model predictions into different terms that are considered as importance scores of the input features. (Schnake et al., 2022; Schwarzenberg et al., 2019; Pope et al., 2019)
- **Surrogate methods** use a simpler and more interpretable surrogate model to approximate the underlying GNN predictions as accurately as possible. (Vu and Thai, 2020; Huang et al., 2020)

While there are several tools for GNN explainability, not all are directly applicable to every GNN-based model. The architecture of XHGP is distinct in its support for generic GNN explainability methods, thanks to its graph architecture from the input. This differentiates it from models like LaneGCN and PGP, which, being lane-centered, do not inherently represent agents as graph nodes from the outset, limiting the straightforward application of such tools.

2.3. Interpretable and explainable motion prediction

The need to understand the importance of road users' interactions was previously highlighted by the authors of the Socially-COnsistent and UndersTandable (SCOUT) Graph Attention Network (GAT) (Carrasco et al., 2021). This was investigated using the Integrated Gradients (Sundararajan et al., 2017) technique and visualization of learned attention. The illustrations provided for four carefully selected scenarios showed the importance of spatial representation of the nearest traffic participants. The use of the learned attention as a mechanism to generate explainable motion predictions was also proposed for transformer-based models (Zhang and Li, 2022).

The explainability of motion prediction was recently explored by use of conditional forecasting (Khandelwal et al., 2020). The authors of *What if Motion Prediction* (WIMP) model (Khandelwal et al., 2020) developed an iterative, graphical attention approach with interpretable geometric (actor-lane) and social (actor–actor) relationships that support the injection of counterfactual geometric targets and social contexts. In this way, the model can make different predictions based on injected/removed lanes and actors. The proposed method supports the study of hypothetical or unlikely scenarios, so-called counterfactuals. This capability can be further used in the planning process to include only relevant features in the calculations or to examine and evaluate the model's predictions.

In Zhang et al. (2022b), the authors investigated the robustness of motion prediction for AV by proposing a new adversarial attack that disrupts normal vehicle trajectories to maximize prediction error. This study focuses on 3 predictive models and 3 trajectory datasets. The evaluation showed that the prediction models are generally susceptible to perturbation by adversaries and can result in unsafe AV behavior, such as harsh braking. The authors highlighted the need to evaluate the scenario of traffic prediction models. Data augmentation and trajectory smoothing were proposed as mitigation methods. This reduced the prediction error under attacks by 28%.

Bahari et al. (2022) conducted an analysis of off-road predictions using state-of-the-art models. Their study demonstrated that four selected multi-modal motion prediction approaches fail to generalize to new scenes. The authors generated realistic adversarial examples that caused the models to predict off-road scenarios. This analysis was performed on the Argoverse Dataset (Chang et al., 2019). The findings underscore the significance of explainability and interpretability of the predictions to enhance model robustness and generalizability. However, we posit that Argoverse dataset, with its limited variability in scenes and road topologies, as well as minimal interactions among agents, may not be the most suitable choice for thoroughly examining explainability of motion prediction systems, especially in the context of real-world urban scenarios.

To our knowledge, while there have been efforts to understand motion prediction in autonomous driving, a comprehensive approach towards explainable graph-based motion prediction, especially for highly interactive and challenging datasets like nuScenes (Caesar et al., 2020), has been limited. XHGP, as presented in this work, bridges this gap by not just offering accurate motion predictions but also ensuring these predictions are interpretable and explainable, distinguishing it from others in the domain.

3. Methodology

We propose a novel approach for motion prediction named Explainable Heterogeneous Graph-based Policy (XHGP). This model extends the lane-graph traversals idea presented in Deo et al. (2021) to jointly model lanes and different types of dynamic agents in a heterogeneous graph (Yang et al., 2021). In addition, we leveraged the interpretability features of the XHGP model to analyze predicted trajectories. To this end, we applied multiple explainable techniques, including learnt attention visualization, learnt GNNExplainer masks, and counterfactual reasoning. We thoroughly evaluated each of these approaches in the context of multi-modal motion prediction.

3.1. Problem formulation

The goal is to estimate future trajectories of vehicles of interest from their past trajectories and scene context. Input features describe the past trajectory of the vehicle of interest as well as the scene context.

One the one hand, let $s_i^t \in \mathbb{R}^6$ be the state of agent i at time t, including its two-dimensional position in the BEV-plane, $\tau_i^t = [x_i, y_i]$, as well as the velocity v_i, acceleration a_i, yaw rate ω_i, and a flag I_i indicating the type of agent — pedestrian ($I_i = 1$) or vehicle ($I_i = 0$). Therefore $s_i^t = [x_i, y_i, v_i, a_i, \omega_i, I_i]$ (see Fig. 2(a)). On the other hand, lane centerlines are represented by a segment of fixed

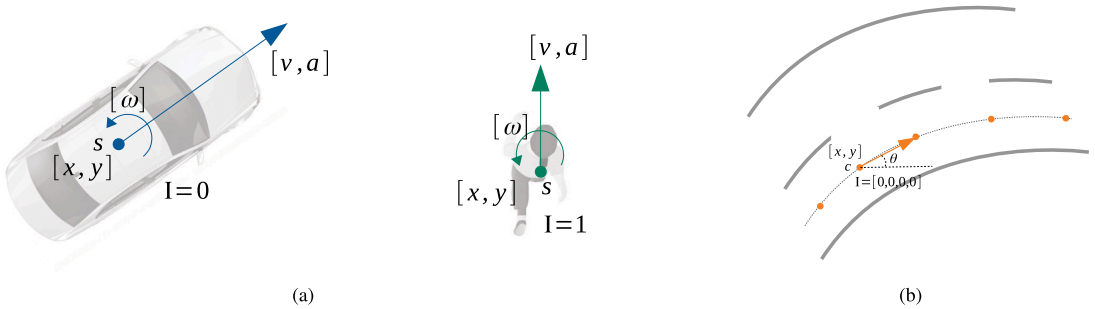

Fig. 2. Illustration of the state variables (a) of an agent and (b) of a lane centerline point.

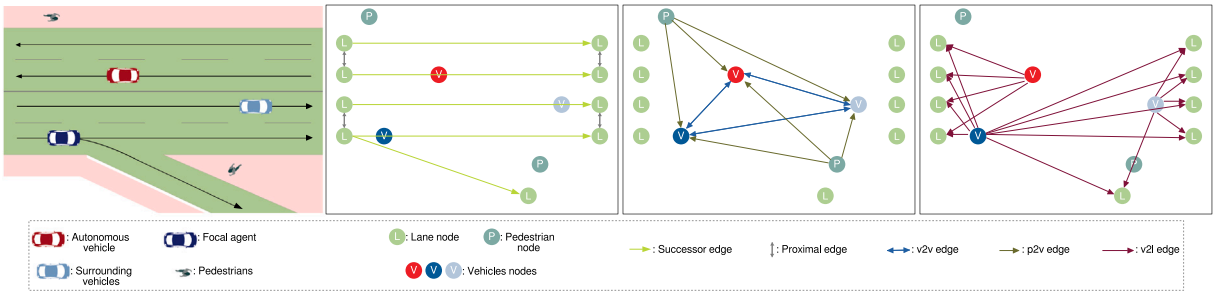

Fig. 3. Example of building a heterogeneous graph from a traffic scene. Nodes include vehicles, pedestrians and lanes. Directed and undirected edges include multiple types of interactions.

length, each segment consisting of a sequence of N points. Each lane point n is defined by a vector $c_n \in \mathbb{R}^7$ which includes its two-dimensional position in the BEV-plane $[x_n, y_n]$, the yaw θ_n, and a semantic 4D binary vector I_n indicating whether the point belongs to a stop line, turn stop, crosswalk, or traffic light. That is, $c_n = [x_n, y_n, \theta_n, I_n]$ (see Fig. 2(b)).

The aim is to predict τ_i^t for future time steps $T_{\text{obs}} < t \leq T_{\text{pred}}$, being $T_{\text{pred}} = 6$ at 2 Hz.

3.2. Multi-modal motion prediction model

Future behaviors of road agents are inherently uncertain and conditioned on the scene context, including road topology and social interactions with other road agents. According to this idea, we explored different model architectures to explicitly capture these two interactions with respect to the agent of interest.

We extend (Carrasco et al., 2021), an attention-based GNN that uses a flexible and generic representation of the scene as a graph for modeling interactions predicting socially-consistent trajectories by adding the scene context as a rasterized HD Map and a multimodal prediction header. However, rasterization has proven to be inefficient and somewhat prone to information loss. Recently, lane-graph based methods have shown better performance and efficiency. On the other hand, our experiments with multi-modal autoencoders based on mixture density networks and on variational autoencoders showed to be prone to mode collapse, with different modes being different in speed profile, landing on a single path. The approach taken in Deo et al. (2021), based on lane-graph traversals conditioning, has shown promising results. They leverage the strong inductive bias that lanes provides to the network, which capture both the direction of traffic flow and the legal routes for each agent in the scene.

In our approach, we model traffic scenarios as heterogeneous graphs, with nodes representing the different agents and static objects in the scene, and edges representing their interactions. This is a more natural and flexible way of encoding the scene context than rasterization, since it can handle different input sizes and achieve invariance to input ordering. Using a generic representation of the scene as a graph allows the model to be flexible to different environments and road topologies and to harness the expressive power of GNNs. Additionally, graph-based representations contain important topology information, which makes them more interpretable than other DL approaches in this scenario, being able to directly retrieve the importance of each interaction in the graph.

Graph definition. We propose to model the whole scene as an heterogeneous graph including both scene and agent context information (see Fig. 3). As stated in Zhang and Li (2022), the use of heterogeneous graphs makes it possible to merge information on road topology and traffic rules, as well as to rationalize the implicit interactive information of the agents. Hence, we define the whole scene context as a directed heterogeneous graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{T}, \mathcal{R}\}$. This heterogeneous graph representation is central to XHGP's design, allowing us to capture interactions and entities in a unified framework. It differentiates XHGP from state-of-the-art graph-based prediction models, which do not capture the full breadth and depth of scenarios in such an integrated manner.

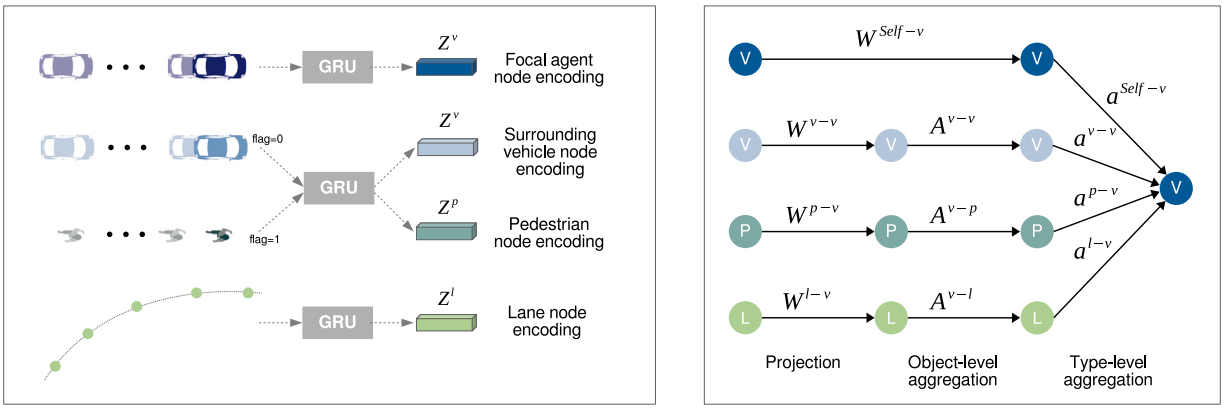

Fig. 4. Left: lane and agent node encoding using different Gate Recurrent Units for the focal agent, surrounding vehicles, and lane nodes. Right: object- and type-level aggregation. W projects the representation of each node to a common semantic space. A is the adjacency matrix used for object-level aggregation. a is the attention mechanism for type-level aggregation.

In this graph, each node $v \in \mathcal{V}$ has a *type*, denoted by a mapping function $\tau(v) : \mathcal{V} \to \mathcal{T}$, and each edge $e \in \mathcal{E}$ has a *type*, represented by a mapping function $\phi(e) : \mathcal{E} \to \mathcal{R}$. The sets $\mathcal{T}$ and $\mathcal{R}$ represent node types and relation types, respectively. The set of node type $\Omega \in \mathcal{T}$ is given by $\mathcal{V}^{\Omega}$, while the neighborhood of Ω is denoted as $\mathcal{N}^{\Omega} = \{\Gamma | \Gamma, \Omega \in \mathcal{T}, \langle \Gamma, \Omega \rangle \in \mathcal{R}\}$, with $\Gamma \in \mathcal{N}^{\Omega}$ as a neighbor type of Ω. The interaction from Γ to Ω is represented as $\langle \Gamma, \Omega \rangle$.

In our work, $\mathcal{T} = \{\text{lane}, \text{vehicle}, \text{pedestrian}\}$, and $\mathcal{R} = \{\text{succ}, \text{prox}, \text{v2l}, \text{v2v}, \text{p2v}\}$. These node types represent lanes, vehicles, and pedestrians, and the edge types define successor, proximal, vehicle-to-lane, vehicle-to-vehicle, and pedestrian-to-vehicle interactions, respectively. We follow the approach described in Deo et al. (2021) to define each lane node, $v \in \mathcal{V}^{\text{lane}}$, as a lane centerline segment of 20 m length. Each node is defined by a sequence of points with its feature vectors $c^l = [c_1^l, \dots, c_N^l]$ as described in 3.1.

Successor edges connect lane nodes along the lane, while proximal edges connect neighboring lane nodes that are within a distance threshold of each other and have a yaw angle difference within a threshold. We consider all agents and lanes within a specific area around the focal agent. Concretely, 50 m laterally, 20 m behind and 80 m in front of the agent. For E^{v2l}, we consider all lane nodes within 10 m around the particular vehicle. The focal agent is connected to all lanes within the predefined region. For E^{v2v} and E^{p2v}, we define an interaction when the euclidean distance between two agents is lower than 20 m.

We construct an adjacency matrix for each relation to represent the connections between nodes of different types. For example, the adjacency matrix for vehicle-to-lane edges would be a matrix of size $|\mathcal{V}^{\text{vehicle}}| \times |\mathcal{V}^{\text{lane}}|$, where the element at row i and column j represents whether the vehicle node i and the lane node j are connected. Fig. 3 provides an illustrative representation of the structure of the graph.

Scene and agent context encoding. Each object type is first encoded using a gated recurrent unit (GRU). We use a separate encoder for the focal agent, surrounding agents – where the type of agent (vehicle or pedestrian) is indicated in the input features –, and lanes nodes. Then, we obtain the encoding Z^{v}, Z^{p}, and Z^{l}.

After encoding and projecting all the hidden representations of neighbor objects into a common semantic space, we employ an attention-based ***object-level*** aggregation similar to that in Carrasco et al. (2021), Veličković et al. (2018), instead of the statistically *row-normalized* convolution operation. For each canonical edge type in $\mathcal{E}$, a subgraph is considered for the aggregation, with source $\mathcal{V}^{\Omega}$, destination $\mathcal{V}^{\Gamma}$ and relation $\mathcal{R}^{\Omega-\Gamma}$. We then compute a pair-wise un-normalized attention score between each two neighbors in the form of additive attention, and normalize them across all the neighborhood of Ω using the softmax function,

$$\alpha_{ij} = \text{Softmax}_j(a^T \cdot \text{LeakyReLU}([W z_i \parallel W z_j])) \ , \tag{1}$$

where z is the node encoding, W is the weight matrix that parametrizes a shared linear transformation applied to every node, a defines the self-attention mechanism, $\cdot^T$ represents transposition, and $||$ is the concatenation operation (see Fig. 4).

The attention score indicates the importance of node j features to node i, which allows each node to attend every other node in its neighborhood. This score is later visualized to understand which interactions are most important. The aggregation of the neighbors embeddings is similar to GCN (Kipf and Welling, 2017), where all the neighbors in the subgraph are aggregated together scaled by the attention score.

To learn more powerful representations for each node type, Yang et al. (2021) proposes ***type-level*** attention to fuse representations from different types of neighbor objects. The intuition is that for each type of object, the information coming for different types of neighbor objects could not have the same importance. For example, for objects of type *lane*, the relation *vehicle-to-lane* could be more relevant for the final prediction than the relation *proximal*. For objects of type *vehicle*, the relation *vehicle-to-vehicle* in general should be more relevant than *pedestrian-to-vehicle*. Hence, before aggregating the representations from the object-level attention aggregation, we first learn the importance for the different types of neighbors using scaled dot product attention (Vaswani et al., 2017). In order to compute the importance of the representations of neighbor types, including the self representation, for

$\mathcal{V}^{\Gamma}$, we linearly project the hidden representation from the previous step, $\{\mathbf{H}^{\Omega-\Gamma}\} \cup \{\mathbf{H}^{\Gamma}\}$, into keys, and $\{\mathbf{H}^{\Gamma}\}$ into the query. The values will be the output from the previous step.

Then, the type-level attention is computed as follows:

$$e^{\Gamma} = \mathrm{ELU}([\mathbf{H}^{\Gamma} \cdot \mathbf{W}_k^{\Gamma} \parallel \mathbf{H}^{\Gamma} \cdot \mathbf{W}_q^{\Gamma}] \cdot w^{\Gamma}) \tag{2}$$

$$e^{\Omega-\Gamma} = \mathrm{ELU}([\mathbf{H}^{\Omega-\Gamma} \cdot \mathbf{W}_k^{\Gamma} \parallel \mathbf{H}^{\Gamma} \cdot \mathbf{W}_q^{\Gamma}] \cdot w^{\Gamma}), \quad \Omega \in \mathcal{N}^{\Gamma} \tag{3}$$

Finally, a softmax function is applied to get the normalized attention coefficients, a_i^{Γ} and $a_i^{\Omega-\Gamma}$. These attention scores are then used to compute the higher level representation of $\mathcal{V}^{\Gamma}$. For each element i in $\mathcal{V}^{\Gamma}$,

$$\hat{\mathbf{H}}_i^{\Gamma} = \mathrm{ELU}(a_i^{\Gamma} \cdot \mathbf{H}_i^{\Gamma} + \sum_{\Omega \in \mathcal{N}^{\Gamma}} a_i^{\Omega-\Gamma} \cdot \mathbf{H}_i^{\Omega-\Gamma}) \tag{4}$$

We apply two layers of the defined GNN to aggregate the two-hop neighborhood information. The node encoding size and the attention size is 32.

Graph traversal and decoding. The final representation for the focal agent, $\hat{\mathbf{H}}_0^v$, is concatenated with its encoding previous the aggregation, Z_0^v. This output, together with the representation for the lane nodes, $\hat{\mathbf{H}}_0^l$, is used to learn a policy for graph traversal using behavior cloning. The final decoding is conditioned on path traversals, following Deo et al. (2021).

3.3. Towards explainable motion prediction

Our main objective is to integrate explainability techniques on the XHGP model, to implicitly reflect the influence of the different agents and the topology of the road on the multimodal predicted trajectories, without negatively impacting the performance of the system. More specifically, we integrate three different explainability techniques in the graph-based model: analysis of attention, post-hoc explainability (GNNExplainer) and counterfactual reasoning.

3.3.1. Attention visualization

According to previous research (Jain and Wallace, 2019), attention should not be seen as a direct indicator of meaningful explanations, but rather as a proxy for explainability. However, in the context of GNNs used to model road scenarios, attention can implicitly capture explicit interactions between agents and road topology. This representation is pivotal for XHGP, enabling it to differentiate interactions more comprehensively than other graph-based architectures, making it particularly suited for our use case.

As explained in Section 3, the scene context encoding is performed using two types of attention: object-level and type-level attention. The former reflects the importance of each neighboring node in the aggregation, i.e., how each interaction affects the prediction. The latter represents the importance of each type of interaction. We aggregate both attention scores and visualize them to understand which interactions in the scene are most relevant for the final prediction. The ability to visualize and understand the importance of interactions, especially with dynamic agents, is an example of the explanatory power of the XHGP design. State-of-the-art models, while effective for motion prediction, do not natively offer this level of granularity in explaining the influence of individual interactions on the predicted outcomes.

3.3.2. Post-hoc explainability methods: GNNExplainer

Existing methods adapting or specifically designing the explanation methods for GNNs have shown promising explanations on multiple types of graph-structured data (Jaume et al., 2020; Rao et al., 2022; Jaume et al., 2021; Zhdanov et al., 2022). In particular, perturbation methods involve learning or optimization (Ying et al., 2019; Luo et al., 2020a; Schlichtkrull et al., 2021; Lin et al., 2021; Lucic et al., 2021). While they come with higher computational costs, they generally achieve state-of-the-art performance in terms of explanation quality. The XHGP architecture allows us to leverage these methods more effectively, particularly in the context of heterogeneous graphs representing intricate driving scenarios. These methods train local post-hoc interpretable models on top of the explained predictive model. One of the most interesting approaches is GNNExplainer (Ying et al., 2019) which requires training or optimizing an individual explainer for each data instance, i.e., a graph or a node to be explained.

GNNExplainer is a graph pruning based explainability technique, which can be used with any type of GNN. The explainer tries to find the minimum sub-graph $G_s \in G$ such that the model prediction is retained. The task is formulated as an optimization problem that learns a mask to activate or deactivate parts of the graph. The initial formulation (Ying et al., 2019) was developed for explaining node classification tasks and is based on edge masking. However, other works (Jaume et al., 2020) utilized the technique also to learn a mask over the nodes instead of edges.

Formally, GNNExplainer aims to maximize mutual information between a compact sub-graph G_s and a base graph G by learning an edge mask and a feature mask in an optimization process (Ying et al., 2019). The undoubted advantage of the methodology is its versatility, that is, its adaptability to problems other than graph or node classification or link prediction, for which it has been used in previous studies. We adapt the approach to work on heterogeneous graphs and learn the importance mask of its edges (traffic agents interactions) for the prediction of the focal agent trajectory for each timestamp in the given scenario. Our objective is to provide a comprehensive explanation of the modeled traffic scene graph. To achieve this, we employ GNNExplainer masking during the stage of encoding scene and agent context information. Subsequently, we proceed with the remaining stages of the model to compute the trajectory for the selected mode. This integration facilitated the extraction of detailed explanations from the GNN-based

model's behavior, providing a sound understanding of the reasoning behind predicted trajectories. We intend the explanations to be as compact as possible, while providing the same prediction as the original graph. Heuristically, we enforce these constraints by optimizing the following objective function (Ying et al., 2019):

$$\mathcal{L} = \mathcal{L}_{dist}(y, \hat{y}) + \alpha_1 \|M\|_1 + \alpha_2 \mathcal{H}(M) + \beta_1 \|F\|_1 + \beta_2 \mathcal{H}(F), \tag{5}$$

where $\mathcal{L}$ is the loss function measuring distance between original and predicted trajectory of interest (usually the most probable first mode), $\hat{y}$ is the original model prediction, y is the model prediction with the edge (M) and feature (F) mask applied. In GNNExplainer (Ying et al., 2019), an entropy-like metric $\mathcal{H}$ is applied by transforming the M and F matrices into a probability-like distribution and then calculating a heuristic measure of information content. The importance masks for edge and node features, denoted as M and F respectively, are continuous and range between 0 and 1. The magnitude of these masks indicates the degree of importance, with larger values indicating higher importance. These masks are associated with the canonical types of edges and nodes in the computation graph. To measure the distance between different predicted trajectories we use the same metric as for the predictive model (Deo et al., 2021; Zhao et al., 2020). In the end we obtained the edge weights, which we were able to compare with the attention learned by the GNN encoder of the XHGP model. Moreover, unlike existing open implementations designed for homograph data, our approach addressed the challenges posed by the diverse nature of the input data in form of a heterogeneous graph representing interactions between both static and dynamic objects in autonomous driving scenes. To further enhance interpretability of the explanations, we developed visualization techniques that showcase the most influential edges (indicating interaction between dynamic objects) within the heterogeneous graph, thereby enabling a clear and intuitive understanding of the GNN-based model's prediction process. In this way, our work pioneers the adaptation, integration, and applicability of GNNExplainer in the context of GNN-based motion prediction for autonomous driving scenarios. Our approach showcases how GNNExplainer can be successfully applied to this domain.

3.3.3. Counterfactuals

Mentioned above methods are not counterfactual by their nature, meaning that they do not explain how to achieve an alternative outcome from the predictive model. Counterfactuals about hypothetical occurrences are increasingly used in different explainable AI applications. A counterfactual explanation describes a cause and effect situation in the form *If X had not occurred, Y would not have occurred*. The name *counterfactual* comes from imagining a hypothetical reality that contradicts the reality. Counterfactual reasoning can be used to explain the predictions of individual instances made by a model by modifying the cause (input features) of that prediction. This makes counterfactuals one of the most intuitive explanation for humans since, first, reasoning about cause and effect in a situation comes very natural to humans, and, second, they focus on a specific instance and a small input change. In our study we presented how changing different part of the modeled traffic scenario can affect the final prediction.

Our method models the entire scene as a heterogeneous graph, including interactions between the road network and dynamic agents. This modeling approach naturally enhances XHGP's capability for counterfactual reasoning. By capturing geometric and social relationships, our model facilitates a more nuanced exploration of hypothetical scenarios. The surrounding context of the focal agent can be manipulated to ablate specific social or road influences, as well as to condition upon unobserved hypothetical agents.

These counterfactual explanations can be used not only to explain, but also to evaluate predictive models that rely on their scene context. Furthermore, they can be seen as a measure of robustness and generalizability. We want our model to produce sensible predictions when conditioned on unlikely extreme inputs, demonstrating that the model has learned a powerful causal representation of driving behavior. This capability is not only useful for explainability, but could also be exploited by a subsequent decision making system or subsequent planner to reason about social influences from occluded regions.

4. Evaluation framework

4.1. Dataset

In our experiments, the reference dataset is the public, large-scale nuScenes dataset (Caesar et al., 2020) for autonomous driving developed by the Motional team. The primary objective behind the creation of this dataset was to enhance the safety, reliability, and accessibility of self-driving cars in everyday life. To achieve this goal, the dataset encompasses a diverse range of scenarios involving interactions between various traffic participants, making it the most suitable choice for our analysis on explainability among other public benchmarks (Chang et al., 2019; Ettinger et al., 2021; Mandal et al., 2020; Zhan et al., 2019; Caesar et al., 2020). The nuScenes dataset consist of 1000 scenes of 20 s each, with ground-truth annotations and HD maps, which were collected in Boston and Singapore, where right-hand and left-hand traffic rules apply respectively. In our prediction scenario, we rely on 2 s of dynamic agents' history and the map features to predict the next 6 s. Given the geographic diversity of the data and comprehensive representation of complex scenarios such as turns and intersections, nuScenes presents one of the most challenging prediction benchmarks, which enables us to thoroughly evaluate the performance and robustness of the model in a multitude of real-world situations, furthering the understanding of its capabilities and limitations. For presented results, the challenge split from the nuScenes motion prediction benchmark (Caesar et al., 2020) was used.

4.2. Evaluation metrics

4.2.1. Quantitative metrics for motion prediction

first, we perform a quantitative evaluation of the motion prediction model using nuScenes motion prediction benchmark. We output 10 predictions for the focal agent, 6 s into the future at 2 Hz, along with the probability that the agent follows that trajectory. The metrics used in the nuScenes benchmark to measure the degree to which this proposed set of trajectories matches the ground-truth are as follows:

- Best-of-K Average Displacement Error (minADE): The minimum point-wise L2 distance to the ground-truth trajectory over all predicted trajectories.
- Best-of-K Miss Rate (MR): percentage of predictions whose maximum pointwise L2 distance between the prediction and ground-truth is greater than 2.0 m.
- Off-road rate: computes the fraction of trajectories that are off-road, outside the drivable area.

4.2.2. Quantitative metrics for graph explainability

In order to evaluate the GNN explanation method, it is necessary to qualitatively examine the results for each input example. This would be extremely time-consuming with only visualizations. In addition, this type of evaluation is highly dependent on people's subjective understanding, resulting in a biased and unfair evaluation. Evaluation metrics are a good alternative to study these explanation methods. In this section, we introduce several recently proposed evaluation metrics (Yuan et al., 2022).

Sparsity. Good explanations should capture the most important input features and ignore the irrelevant ones. The metric *Sparsity* measures the fraction of features selected as important by the explanation method (Pope et al., 2019). Formally, for given ith input of the graph G_i and its importance weights m_i, the *Sparsity* metric can be computed as

$$Sparsity = \frac{1}{N}\sum_{i=1}^{N} 1 - \frac{|m_i|}{|M_i|}, \tag{6}$$

where N is the number of graphs, $|m_i|$ denotes the number of important features (i.e., nodes, edges, node features) identified as m_i, and $|M_i|$ — the total number of features in G_i. Higher values of the metric indicate that explanations are more sparse and tend to capture only the most important input information.

Fidelity. The explanations provided should be faithful to the model, meaning that they should identify input features that are important to the model, not (only) to the human. To evaluate this, the *Fidelity+* (Pope et al., 2019) and *Fidelity−* (Yuan et al., 2022) scores were introduced. The *Fidelity+* is defined as the difference of accuracy (or predicted probability) between the original predictions and the new predictions after masking out important input features. For the given $i-$th input graph G_i, the score can be computed as

$$Fidelity+ = \frac{1}{N}\sum_{i=1}^{N}\Big(1 - \mathbb{1}(y_i^{1-m_i} \simeq y_i)\Big), \tag{7}$$

where y_i is the original prediction of graph i. $1 - m_i$ means the complementary mask that removes the important input features, and $y_i^{1-m_i}$ is the prediction when feeding the new graph into the trained GNN-based $f(\cdot)$. The indicator function $\mathbb{1}(y_i^{1-m_i} \simeq y_i)$ returns 1 if $y_i^{1-m_i}$ and y_i are with an Euclidean distance lower than 2 m, and returns 0 otherwise. For GNNExplainer (Ying et al., 2019), the importance scores are continuous values, and hence the importance map m_i can be obtained by normalization and thresholding or ranking (Yuan et al., 2022). For *Fidelity+* higher values indicate better explanation results and more discriminative features are identified.

In contrast, the metric *Fidelity−* studies prediction change by keeping important input features and removing unimportant features, and is defined as

$$Fidelity- = \frac{1}{N}\sum_{i=1}^{N}\big(1 - \mathbb{1}(y_i^{m_i} \simeq y_i)\big), \tag{8}$$

where $y_i^{m_i}$ is a new predicate based on the explanation of m_i for $G_i^{m_i}$ being a new graph by preserving only the important features of G_i. For *Fidelity−*, lower values indicate less importance information are removed so that the explanations results are better.

Note that, when the explanation results are soft values (i.e., $m_i \in [0, 1]$), as it is in our case, the *Sparsity is determined by a threshold value. Intuitively, larger threshold values tend to identify fewer features as relevant and, hence, increase the Sparsity score and decrease the Fidelity+ score.*

5. Results

5.1. Quantitative evaluation of multi-modal motion prediction

We compare our method against state-of-the-art models on nuScenes benchmark. Results are shown in Table 1. This method achieves similar results as the original PGP. The results described for PGP are obtained by reproducing their work, which improve on those presented in the original paper. In the table, we detailed both the results presented in their work and those obtained

Table 1
Comparison to the state-of-the-art on nuScenes benchmark. We change the PGP encoding, modeling the entire scene as a heterograph, to make it compatible with GNN explanability methods without any loss of performance.

Method	MinADE		Miss rate		Off-road
	K = 5	K = 10	K = 5	K = 10	
Trajectron++ (Salzmann et al., 2020)	1.88	1.51	0.70	0.57	0.25
WIMP (Khandelwal et al., 2020)	1.84	1.11	0.55	0.43	0.04
CXX (Luo et al., 2020b)	1.63	1.29	0.69	0.60	0.08
P2T (Deo and Trivedi, 2020)	1.45	1.16	0.64	0.46	0.03
THOMAS (Gilles et al., 2022)	1.33	1.04	0.55	0.42	0.03
PGP (Deo et al., 2021)	1.30	1.00	0.61	0.37	0.03
	1.28†	0.95†	0.52†	0.34†	0.03†
XHGP	1.27	0.94	0.53	0.34	0.03

with the new implementation ($\cdot^{\dagger}$). Both these models achieve best results in nuScenes motion prediction leaderboard (Caesar et al., 2020). While the performance metrics are comparable, it is important to note that XHGP's architecture is inherently designed for explainability. The true strength of XHGP lies in its ability to provide meaningful, interpretable insights into its predictions, a feature that distinguishes it from other models like PGP.

Fig. 5 shows an example predicted scene. Two seconds of history are represented by a dashed black line. Six seconds of future by a white dashed line. The AV is shown in red and the focal agent in dark blue. Surrounded vehicles are represented in light blue, bicycles and motorcycles in green dots, and pedestrians in blue dots. Movable objects, such as traffic cones, are painted in yellow. Ten different predictions – modes – are visible in the scene. The probability of each prediction is assigned to the color bar on the right. Therefore, each prediction is represented with a color and a line width describing its probability. In this sample scene, the most probable prediction goes straight forward. Two less likely modes turn right. The rest of the variability lies in the velocity profile.

5.2. Evaluation of the influence of interactions

The black-box nature of neural networks raises many concerns, as the predictions they provide are difficult for humans to fully understand. The need of the explanations in autonomous driving stems from existing problems, established regulations and standards, along with opinions of the public. The XHGP model, with its heterogeneous graph representation, is specifically designed to tackle this challenge. By being able to pinpoint influential interactions and provide an understanding of the graph's temporal evolution, it offers a level of transparency and insight not commonly found in other state-of-the-art models.

Since our graphs modeling the scene context are not very large and contain up to 100 nodes, we did not limit ourselves to just the neighborhood of the focal agent. Analysis conducted for the validation subset of the nuScenes dataset identified lanes nodes as the most influential. On the other hand, the most influential type of edges for predicting the most probable trajectory appears to be the ones connecting pedestrian-to-vehicle, and vehicle-to-vehicle. This finding is consistent with one would expect from a human driver, as the ability to perform most maneuvers on the road depends on the geometry of the scenario (lanes, traffic lights, signs, etc.), and their exact performance is based on interactions between dynamic objects (pedestrians, cyclists, and other vehicles). However, we also detected a strong focus on vehicles and pedestrians who do not fully participate in the driven scenario, but merely indicate the limits of the road (sidewalk). This attention may be associated with the fact that, in our graphical representation of the scene, the lanes of the road are defined only by the centerlines.

The influence of interactions was first visualized for the most likely trajectory predicted by our model. To do so, we focus on selected scenes, trying to understand the basic rules for these predictions. Given the findings stated in the previous paragraph, we focus on the interactions with dynamic agents. Fig. 6 depicts explanations, i.e., interaction importance, for a city center scenario with a large number of pedestrians and vehicles — both parked and in movement. The focal agent continues its trajectory in a straight line entering an intersection with the intention of turning right, stopping at the pedestrian crossing. The Figure presents the most important vehicle-to-vehicle and pedestrian-to-vehicle interactions for the focal agent. Top row indicates the most probable predicted trajectory, while the followings rows show the masks learnt by GNNExplainer and the learnt attention weight (marked in red — interactions with vehicles, and in yellow — interactions with pedestrians) between each node connected with the focal agent. Each column represents a different timestamp for the given scenario, moving from left to right. The thickness of the edges represents the importance level of each interaction, which changes for each frame. Overall, GNNExplainer and the visualized attention indicate similar interactions as the important ones, however GNNExplainer appears to be more sensitive for all interactions with the focal agent.

The results, explained in terms of interaction importance, can be produced for different modalities, not only the most probable trajectory. Fig. 7 shows a scenario with the focal agent on a T-junction, attempting to turn right. The first image shows the ten possible modalities predicted by the model, the next images show the importance of interactions between agents for the selected predicted trajectory. We can observe two different behaviors: turning right (most likely prediction, coinciding with the ground-truth), and continuing straight (lower probability). In the former, the interaction between the central agent and pedestrians crossing the road was identified as the most important interaction. In this situation, the car slowly entered the intersection and turned right. For

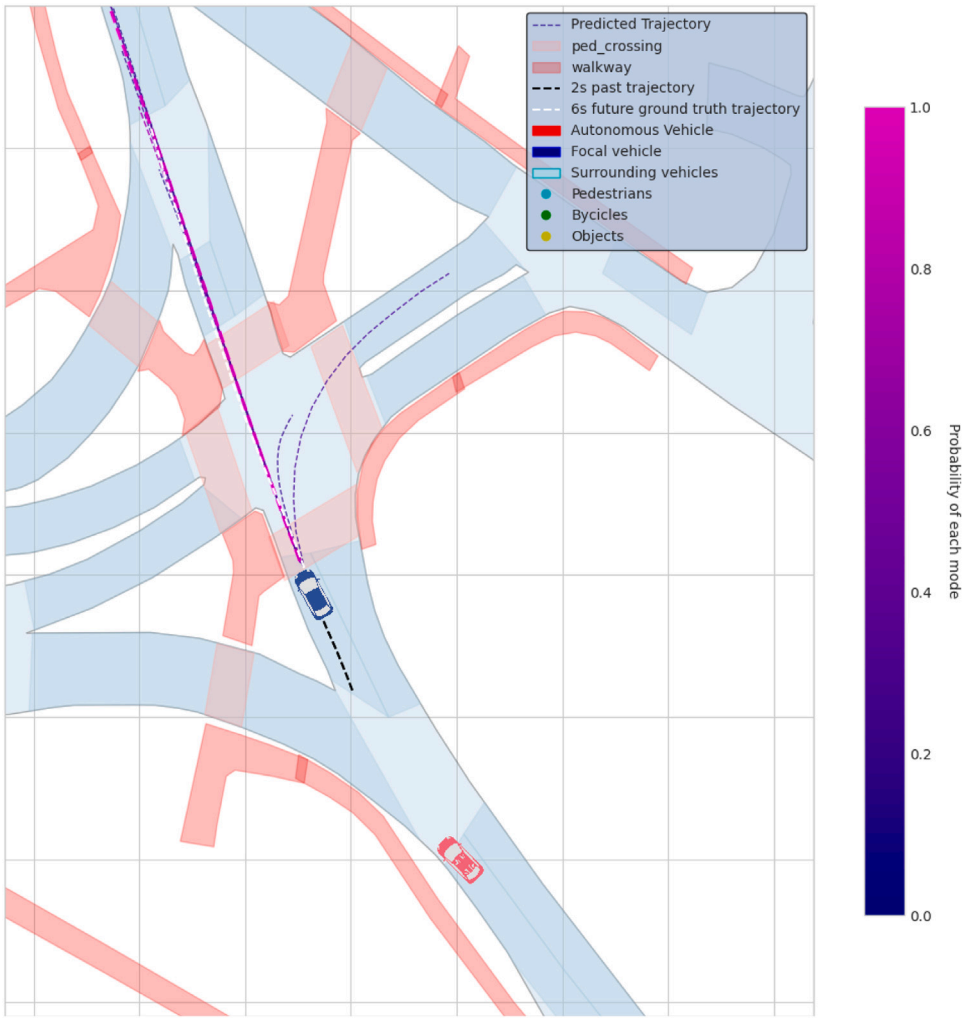

Fig. 5. Multi-modal motion prediction output representation with visual components and legend applied to all presented visualizations.

Table 2
Performance evaluation of GNNExplainer and attention baseline explainability approaches.

Method	*Fidelity+*	*Fidelity−*	*Sparsity*
GNNExplainer	0.47	0.97	0.81
Attention	0.57	0.96	0.64

the other mode, a straight trajectory is observed with an increase in the importance of the interactions with surrounding cars. In this case, the interaction with pedestrians on the sidewalk – which indicates the roadway boundary – was of less importance.

To provide a quantitative validation of the trained explainers for the most probable trajectory, we further calculated metrics described in Section 4.2.2. The Table 2 presents results for GNNExplainer and learnt attentions for $m_i \geq 0.8$ threshold. The scores obtained are similar for both methods, which can also be observed in the qualitative analyses presented above. The higher *Sparsity* value for GNNExplainer indicates that the explanations provided by this method are more sparse — i.e., it tends to give higher importance for fewer input features. This is also related to the main idea of this approach of finding the minimum sub-graph.

5.3. Counterfactual explanations

In this section we explore a series of counterfactual experiments to explain challenging scenarios and demonstrate how the output distribution is highly influenced by the scene context.

Agent insertion. Fig. 9 shows three scenarios where fictitious agents are inserted in the scene or behaviors of current agents are modified.

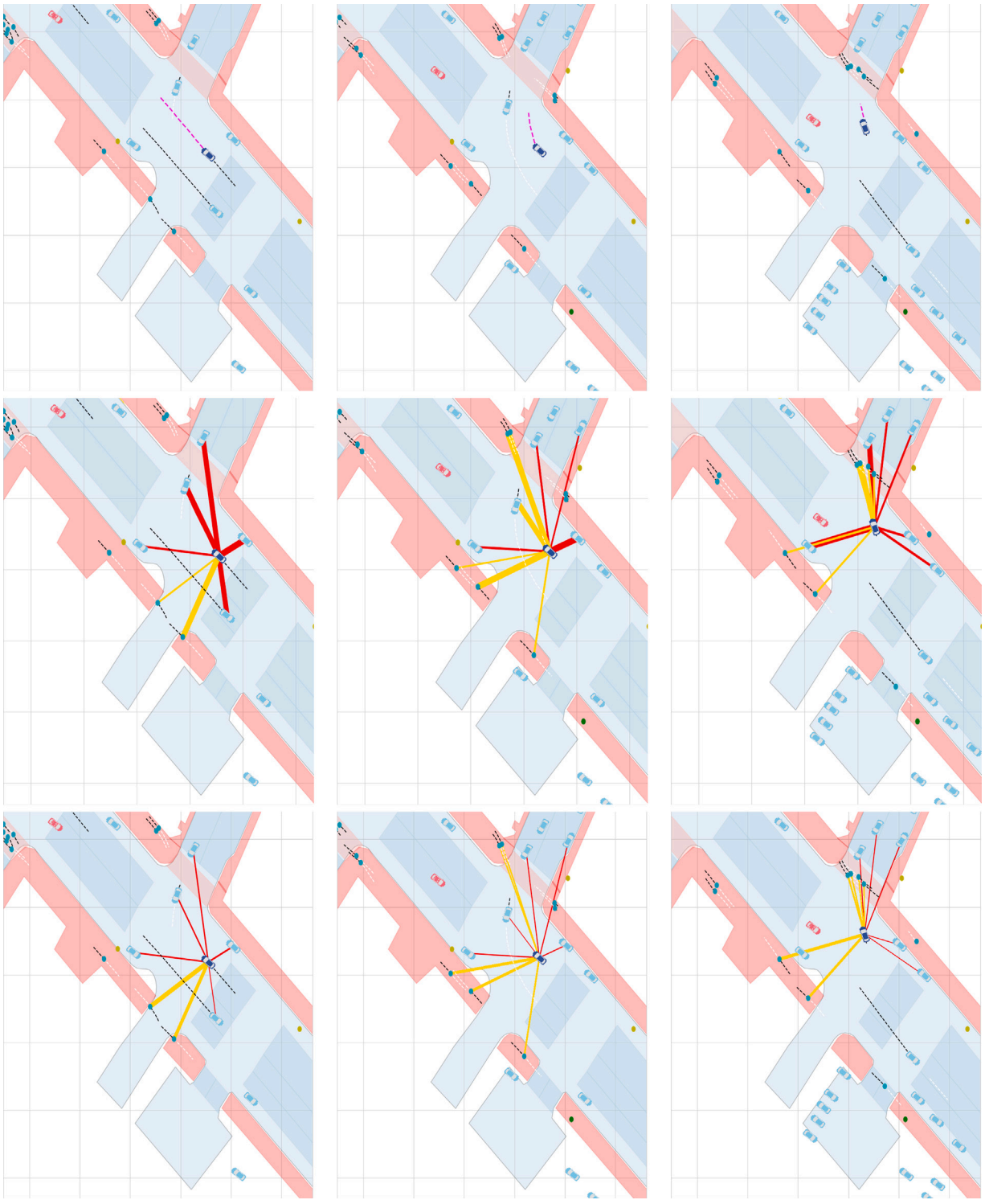

Fig. 6. Qualitative results: importance of interactions for vehicle-to-vehicle (red) and pedestrian-to-vehicle (yellow) with the focal agent based on GNNExplainer (second row) and learned attention (third row). First row depicts the most probable trajectories, which are explained.

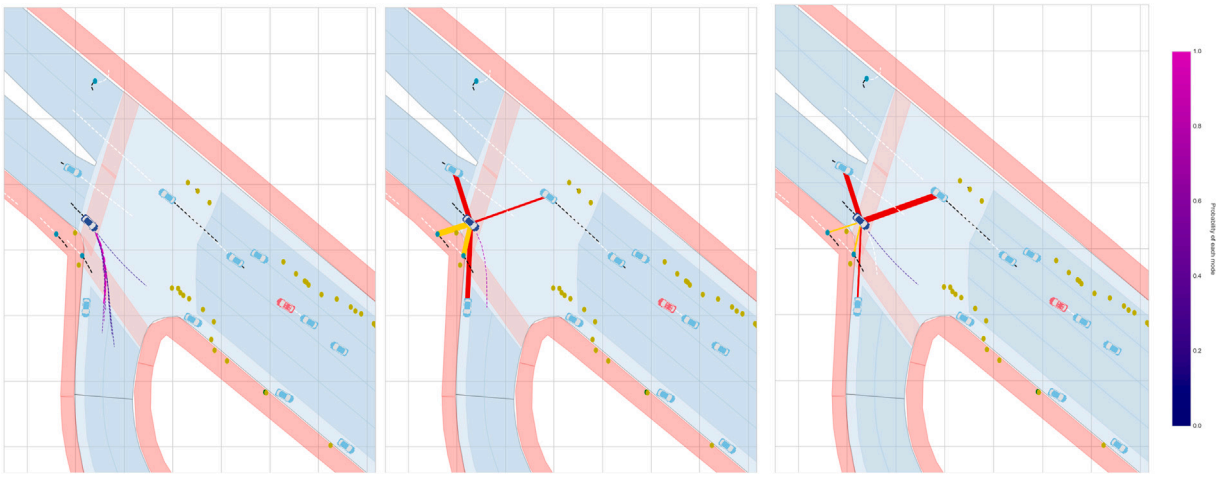

Fig. 7. Qualitative results: importance of interactions for vehicle-to-vehicle (red) and pedestrian-to-vehicle (yellow) with the focal agent based on GNNExplainer for different modalities. The first image shows the 10 possible modes predicted by the model indicating 2 different maneuvers – turning right and going straight – explained in the next images.

Table 3
Ablation of lanes and dynamic agents. Results for K = 10 simulating different recall levels for lane and dynamic agent detection.

Object	Experiment	minADE_5	minADE_10	miss rate_5	miss rate_10	BC
Lanes	Baseline recall 100%, K = 10	1.27	0.94	0.53	0.34	1.85
	Recall 95%	1.36	1.00	0.54	0.36	2.18
	Recall 90%	1.44	1.05	0.57	0.39	2.47
	Recall 80%	1.61	1.16	0.61	0.43	3.08
	Recall 50%	2.11	1.51	0.70	0.55	3.98
	Recall 20%	2.49	1.83	0.76	0.62	4.07
	Recall 0%	2.72	2.11	0.80	0.64	3.86
Dynamic agents	Recall 95%	1.30	0.97	0.52	0.34	1.93
	Recall 90%	1.31	0.97	0.53	0.35	1.93
	Recall 80%	1.32	0.97	0.53	0.34	1.94
	Recall 50%	1.34	0.98	0.54	0.36	1.96
	Recall 20%	1.38	0.99	0.57	0.37	1.91
	Recall 0%	1.38	0.99	0.58	0.37	1.96

In the first scene, the trajectory of the bicycle is modified so that it intersects with the focal agent ground-truth trajectory. The vehicle in front of the focal agent is removed. Three predictions appear turning right, whereas the other seven predictions reduce drastically their velocity to allow the bicycle to safely cross the intersection. If the vehicle in front is not removed from the scene, the predictions appear to be less conservative and follow the vehicle trajectory. However, velocity of the predictions are still reduced and two modes appear turning right.

In the second scenario, we insert a fictitious vehicle stopped in front of the focal agent. In this case, most of the predictions change, turning right instead of following a straight path. However, one prediction with low probability go off the road. The most probable prediction continues straight with a highly reduced speed. We believe that this mode attempts to cover the scenario in which the stopped vehicle continues its trajectory at some point within the six-second window. The two low probability predictions that continue straight should be taken into account with caution in order to plan a safe maneuver.

In the last scenario, a vehicle is inserted into the center of the roundabout. The speed of the most probable mode is significantly reduced to accomodate this vehicle. Yet, some low-probability modes appear to attempt overtaking manouvers. The mode that initially turned left in the roundabout vanishes, which is a reasonable response considering there is no space to turn left due to presence of the counterfactual agent. Conversely, a new mode emerges that opts for a right turn, indicating the model's adaptability in responding to the altered scene.

In conclusion, our counterfactual analysis demonstrates that the model is capable of generating reasonable predictions when faced with unexpected scenarios, thus highlighting its adaptability and usefulness. However, it is crucial to consider that the model is optimized to seek diversity in its output distribution and, as such, may yield low-probability modes that exhibit less cautious or even reckless behaviors, encountering some predictions that may not align well with the altered scenarios. As such, it is imperative for the decision-making system within autonomous vehicles to carefully weigh the likelihood of all predicted scenarios in order to plan trajectories that optimize safety and efficiency. Additionally, it underscores the importance of training models on more complex datasets that cover a wide range of corner-cases to improve their robustness and reliability. By incorporating this understanding of

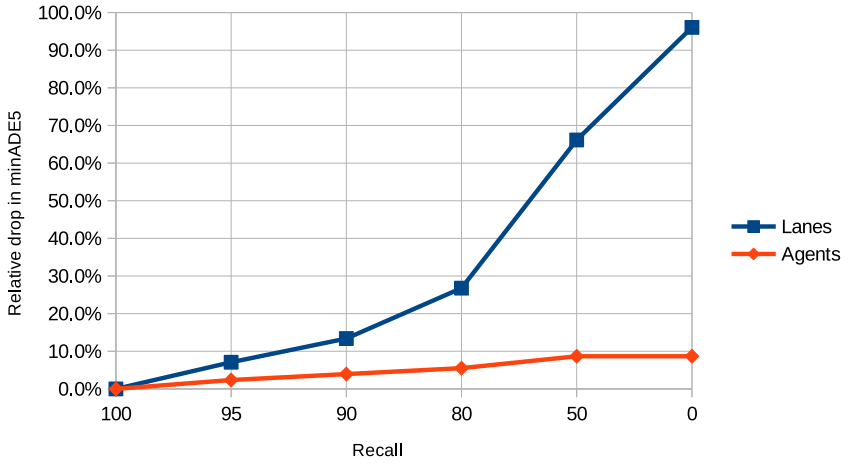

Fig. 8. Relative drop in $minADE_5$ under different recall levels for lane and dynamic agents detection.

the model's strengths and limitations, we can further advance the development of autonomous driving systems that are well-equipped to navigate complex real-world environments.

Object ablation. In addition, we perform a qualitative and quantitative evaluation of the effect of lanes and dynamic agents on the output. Lanes insert important inductive biases in the model and are, hence, crucial for its performance. On the other hand, the ability to capture interactions is essential for a safe and efficient planning. We perform an ablation in an independent manner for lanes and dynamic agents, simulating noise in the perception system. Table 3 shows the performance of the model as having an object detector with a recall from 95% – which is close to the demand for self-driving cars – to 0%, i.e. no lanes or dynamic agents detected at all. Fig. 8 shows the relative evolution of $minADE_5$ in a graphical manner.

When the recall is 0% for lane detection, the system can only rely on the focal agent's past trajectory and other dynamic agents, with no information about the road topology. In case of masking all agents in the scene, we evaluate the impact on performance when interactions are not taken into account for the prediction. For the ablation of dynamic agents, the noise is introduced in the temporal dimension, meaning that we mask random frames of the agents that interact with the focal agent.

Performance is noticeably reduced as the recall of the object detector decreases. The most detrimental effect is shown for lane masking. These results suggest that the model relies heavily on lane information to make its predictions given the significant inductive biases they introduce. However, the results are still sensible when no lanes are detected. Interactions also prove to be an important factor in model performance, albeit in a much lower scale. This is probably due to the fact that most driving trajectories can be inferred solely from the agent's past trajectory and road topology. In addition, the data do not include enough scenes in which interactions are crucial for the behavior of the focal agent. Nevertheless, information from other dynamic agents in the scene is essential to produce more sensible predictions, avoiding collisions and dangerous behaviors that are not consistent with the traffic scene.

In order to gain a deeper understanding of the system's behavior under noisy conditions, we conducted a qualitative evaluation by ablating specific, important lanes. More concretely, we masked lanes traversed by the ground-truth trajectory. For all experiments, we assessed the system's behavior when only some lane segments in the trajectory were masked, as well as when masking the entire ground-truth path. When only one lane segment was removed from the input, the model's behavior remained virtually unchanged. However, when multiple lane segments were masked, the predictions adapted to identify alternative plausible routes.

Fig. 10 presents four sample scenes, with masked lanes indicated by faded red circles. In the first scenario, the probability of the alternative path – continuing straight – increases, with all modes going in that direction. In the second scenario, when the entire ground-truth path is masked, the speed in most predictions is drastically reduced. It appears that the predictions struggle to capture the turn, as they lack the necessary information about the road topology. In the third scenario, there is no other plausible path apart from the ground-truth. Although the predictions change slightly, they still comply to the traffic rules despite the missing information. Finally, in more complex scenarios, like the fourth one, where multiple plausible paths have masked lanes, the model tends to choose a more conservative approach, such as turning right at the intersection. However, when the model does attempt to follow the merging lane, it fails to accurately capture the curvature of the lane.

Overall, these observations demonstrate the model's ability to adapt to various situations, even when faced with incomplete information about the environment. However, the results also emphasize the importance of accurate and comprehensive road topology data for generating more reliable and sensible predictions. As such, future research should continue to focus on improving the model's robustness and performance, particularly under conditions with varying levels of road topology data availability, to ensure the safe and efficient operation of autonomous vehicles in real-world scenarios.

These ablation studies further showcase the power of XHGP's explainability features. The model's ability to adapt to missing or altered information is not just about performance metrics but also about understanding the underlying reasons for the changes in predictions. This level of insight is a distinctive feature of XHGP, setting it apart from other models in the field.

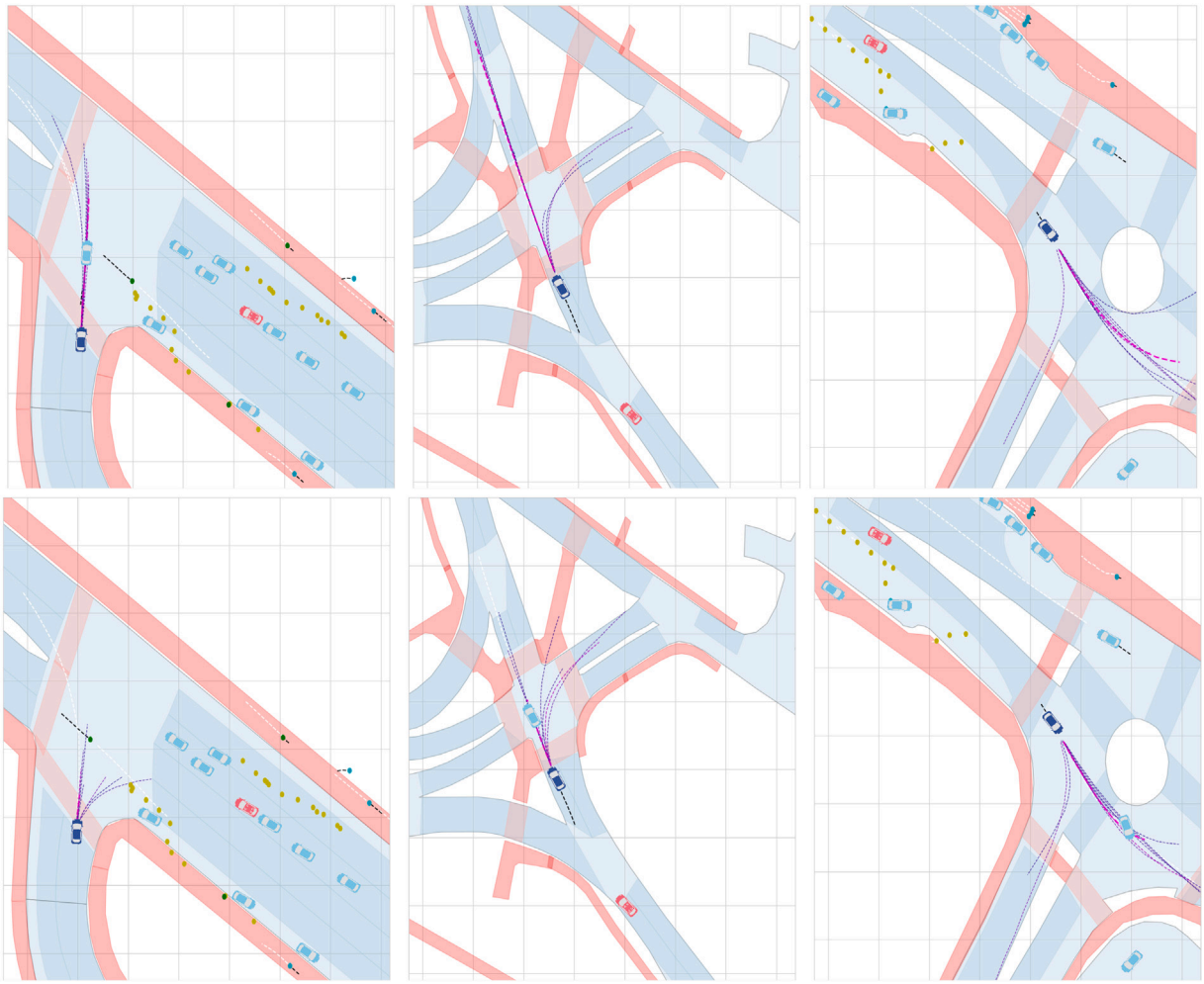

Fig. 9. Qualitative results of counterfactual insertion. Top row shows the real scene with ten predicted modes. Bottom row shows same scene with the counterfactual and the modified predictions. In the first scenario, velocity is highly reduced to avoid the collision with the bicycle, which now intersects the ground-truth trajectory. In the second scenario, the most likely predictions turn right to avoid the vehicle stopped in front. In the last scenario, the velocity of the main mode is highly reduced, the mode turning left disappears due to space constraints, and a new mode opting for a right turn emerges.

6. Discussion and conclusions

In this work, we advance toward explainable motion prediction by proposing a novel approach — Explainable Heterogeneous Graph-based Policy (XHGP). XHGP builds upon previous works by incorporating an heterogeneous graph representation, object-level and type-level attention mechanisms, and lane-graph traversals (Deo et al., 2021). Our model explicitly captures the interactions among all agents and objects in the scene, a key aspect that has not been fully explored in prior works, and enhances the explainability of motion predictions. These unique features of XHGP set it apart from existing approaches and contribute to the advancement of the field.

While state-of-the-art models excel in motion prediction, XHGP's architecture inherently supports a broader spectrum of explainability techniques. We harness tools like GNNExplainer (Ying et al., 2019) and employ counterfactual reasoning, due to its inherent graph architecture from the input, allowing for deeper insights into node and edge interactions.

Our contribution extends beyond merely building on existing methods; we specially tailored the integration and applicability of explainability techniques for multimodal motion prediction in autonomous driving. These adaptations allowed us to interpret the GNN-based model's predictions effectively, shedding light on the decision-making process and fostering trust in the model's capabilities. Visual representations further amplify its interpretability, instilling confidence in safety-critical applications. The combination of these approaches in our analysis serves as a foundation for more transparent and trustworthy motion prediction systems. This is crucial not only for end-users, but also for developers troubleshooting the system, and for type approval and auditing procedures.

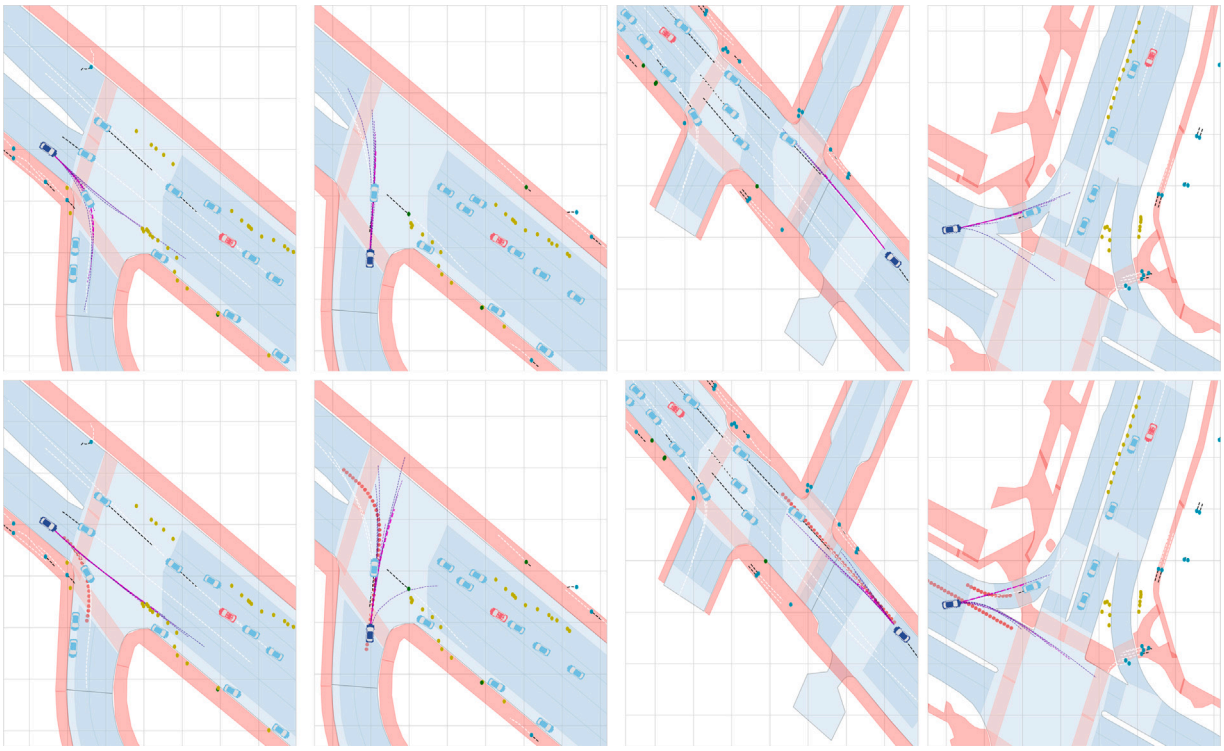

Fig. 10. Qualitative assessment of lane ablation. Top row shows the predictions for 100% recall. Bottom row shows the same scene where some critical lanes have been masked. Masked lanes are represented with faded red circles. In the first scene, all predictions take the alternative path. In the second scenario, predictions fail to capture the turn since the system cannot rely on lane information. In the third scenario, as there is no alternative route, predictions stay robust with slight changes. In the last scenario, most predictions change trajectory, turning right.

As the use of GNNs expands across various fields, the development of explainability techniques continues to grow. Future research aims to leverage these emerging techniques to further improve the explainability of GNN-based motion prediction systems, fostering greater trust and transparency in human–vehicle interactions.

CRediT authorship contribution statement

Sandra Carrasco Limeros: Conceptualization, Methodology, Software, Validation, Investigation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Sylwia Majchrowska:** Methodology, Software, Validation, Formal analysis, Writing – original draft, Writing – review & editing. **Joakim Johnander:** Conceptualization, Writing – review & editing, Supervision. **Christoffer Petersson:** Conceptualization, Writing – review & editing, Supervision. **David Fernández Llorca:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration; The information and views expressed in this paper are purely those of the authors and do not necessarily reflect an official position of the European Commission.

Acknowledgments

This work was supported by HUMAINT project at the Algorithmic Transparency Unit at the Directorate-General Joint Research Centre (JRC) of the European Commission, and in part by the Spanish Ministry of Science and Innovation under Grants PID2020-114924RB-I00 and PDC2021-121324-I00. In addition, part of this work has been carried out during the *Eye for AI Program* thanks to the support of Zenseact and AI Sweden.

References

Bahari, Mohammadhossein, Nejjar, Ismail, Alahi, Alexandre, 2021. Injecting knowledge in data-driven vehicle trajectory predictors. Transp. Res. C 128, 103010.

Bahari, Mohammadhossein, Saadatnejad, Saeed, Rahimi, Ahmad, Shaverdikondori, Mohammad, Shahidzadeh, Amir-Hossein, Moosavi-Dezfooli, Seyed-Mohsen, Alahi, Alexandre, 2022. Vehicle trajectory prediction works, but not everywhere. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR.

Baldassarre, Federico, Azizpour, Hossein, 2019. Explainability techniques for graph convolutional networks. arXiv:1905.13686.

Caesar, Holger, Bankiti, Varun, Lang, Alex H., Vora, Sourabh, Liong, Venice Erin, Xu, Qiang, Krishnan, Anush, Pan, Yu, Baldan, Giancarlo, Beijbom, Oscar, 2020. nuScenes: A multimodal dataset for autonomous driving. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 11621–11631.

Carrasco, S., Llorca, D. Fernández, Sotelo, M. A., 2021. SCOUT: Socially-consistent and UndersTandable graph attention network for trajectory prediction of vehicles and VRUs. In: 2021 IEEE Intelligent Vehicles Symposium, IV. pp. 1501–1508.

Carrasco Limeros, Sandra, Majchrowska, Sylwia, Johnander, Joakim, Petersson, Christoffer, Sotelo, Miguel Ángel, Llorca, David Fernández, 2023. Towards trustworthy multi-modal motion prediction: Holistic evaluation and interpretability of outputs. CAAI Trans. Intell. Technol. 1–16.

Casas, Sergio, Gulino, Cole, Liao, Renjie, Urtasun, Raquel, 2020. SpAGNN: Spatially-aware graph neural networks for relational behavior forecasting from sensor data. In: IEEE International Conference on Robotics and Automation, ICRA. pp. 9491–9497.

Chang, Ming-Fang, Lambert, John, Sangkloy, Patsorn, Singh, Jagjeet, Bak, Slawomir, Hartnett, Andrew, Wang, De, Carr, Peter, Lucey, Simon, Ramanan, Deva, Hays, James, 2019. Argoverse: 3D tracking and forecasting with rich maps. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 8748–8749.

Deo, Nachiket, Trivedi, Mohan, 2020. Trajectory forecasts in unknown environments conditioned on grid-based plans. arXiv:2001.00735.

Deo, Nachiket, Wolff, Eric, Beijbom, Oscar, 2021. Multimodal trajectory prediction conditioned on Lane-graph traversals. In: 5th Annual Conference on Robot Learning. URL https://openreview.net/forum?id=hu7b7MPCqiC.

Doshi-Velez, Finale, Kim, Been, 2017. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

Doshi-Velez, Finale, Kortz, Mason, Budish, Ryan, Bavitz, Chris, Gershman, Sam, O'Brien, David, Scott, Kate, Schieber, Stuart, Waldo, James, Weinberger, David, Weller, Adrian, Wood, Alexandra, 2017. Accountability of AI under the law: The role of explanation. arXiv:1711.01134.

EC, 2021. Regulation of the European parliament and the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. COM(2021) 206 final, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

Ettinger, Scott, Cheng, Shuyang, Caine, Benjamin, Liu, Chenxi, Zhao, Hang, Pradhan, Sabeek, Chai, Yuning, Sapp, Ben, Qi, Charles R., Zhou, Yin, Yang, Zoey, Chouard, Aur'elien, Sun, Pei, Ngiam, Jiquan, Vasudevan, Vijay, McCauley, Alexander, Shlens, Jonathon, Anguelov, Dragomir, 2021. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In: IEEE/CVF International Conference on Computer Vision, ICCV. pp. 9710–9719.

European Commission and Directorate-General for Communications Networks, Content and Technology, 2019. Ethics Guidelines for Trustworthy AI. Publications Office.

Fernández Llorca, D., Gómez, E., 2023. Trustworthy artificial intelligence requirements in the autonomous driving domain. Computer 56, 29–39.

Fernández Llorca, D., Gómez Gutierrez, E., 2021. Trustworthy Autonomous Vehicles. EUR 30942 EN, JRC127051, Publications Office of the European Union, Luxembourg.

Gao, Jiyang, Sun, Chen, Zhao, Hang, Shen, Yi, Anguelov, Dragomir, Li, Congcong, Schmid, Cordelia, 2020. VectorNet: Encoding HD maps and agent dynamics from vectorized representation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 11522–11530.

Gilles, Thomas, Sabatini, Stefano, Tsishkou, Dzmitry, Stanciulescu, Bogdan, Moutarde, Fabien, 2022. THOMAS: Trajectory heatmap output with learned multi-agent sampling. In: International Conference on Learning Representations, ICLR.

Gilpin, Leilani H., Bau, David, Yuan, Ben Z., Bajwa, Ayesha, Specter, Michael, Kagal, Lalana, 2018. Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA. pp. 80–89.

Gonzalo, Rubén Izquierdo, Maldonado, Carlota Salinas, Ruiz, Javier Alonso, Alonso, Ignacio Parra, Llorca, David Fernández, Sotelo, Miguel Ángel, 2022. Testing predictive automated driving systems: Lessons learned and future recommendations. IEEE Intell. Transp. Syst. Mag. 14 (6), 77–93.

Gu, Junru, Sun, Chen, Zhao, Hang, 2021. DenseTNT: End-to-end trajectory prediction from dense goal sets. In: 2021 IEEE/CVF International Conference on Computer Vision, ICCV. pp. 15283–15292.

Gunning, D., 2016. Explainable Artificial Intelligence (XAI). Defense Advanced Research Projects Agency, DARPA-BAA-16-53.

Huang, Yanjun, Du, Jiatong, Yang, Ziru, Zhou, Zewei, Zhang, Lin, Chen, Hong, 2022. A survey on trajectory-prediction methods for autonomous driving. IEEE Trans. Intell. Veh. 7 (3), 652–674.

Huang, Qiang, Yamada, Makoto, Tian, Yuan, Singh, Dinesh, Yin, Dawei, Chang, Yi, 2020. GraphLIME: Local interpretable model explanations for graph neural networks. arXiv:2001.06216.

Jain, Sarthak, Wallace, Byron C., 2019. Attention is not explanation. In: North American Chapter of the Association for Computational Linguistics.

Jaume, Guillaume, Pati, Pushpak, Bozorgtabar, Behzad, Foncubierta, Antonio, Anniciello, Anna Maria, Feroce, Florinda, Rau, Tilman, Thiran, Jean-Philippe, Gabrani, Maria, Goksel, Orcun, 2021. Quantifying explainers of graph neural networks in computational pathology. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 8102–8112.

Jaume, Guillaume, Pati, Pushpak, Foncubierta-Rodríguez, Antonio, Feroce, Florinda, Scognamiglio, G., Anniciello, Anna Maria, Thiran, Jean-Philippe, Goksel, Orcun, Gabrani, Maria, 2020. Towards explainable graph representations in digital pathology. In: ICML 2020 Workshop on Computational Biology, WCB.

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

Kipf, Thomas N., Welling, Max, 2017. Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations, ICLR.

Li, Xin, Ying, Xiaowen, Chuah, Mooi Choo, 2019. GRIP: Graph-based interaction-aware trajectory prediction. In: 2019 IEEE Intelligent Transportation Systems Conference, ITSC. pp. 3960–3966.

Liang, Ming, Yang, Bin, Hu, Rui, Chen, Yun, Liao, Renjie, Feng, Song, Urtasun, Raquel, 2020. Learning lane graph representations for motion forecasting. In: European Conference on Computer Vision, ECCV.

Lin, Wanyu, Lan, Hao, Li, Baochun, 2021. Generative causal explanations for graph neural networks. In: 38th International Conference on Machine Learning.

Lucic, Ana, ter Hoeve, Maartje, Tolomei, Gabriele, de Rijke, Maarten, Silvestri, Fabrizio, 2021. CF-GNNExplainer: Counterfactual explanations for graph neural networks. In: 25th International Conference on Artificial Intelligence and Statistics.

Luo, Dongsheng, Cheng, Wei, Xu, Dongkuan, Yu, Wenchao, Zong, Bo, Chen, Haifeng, Zhang, Xiang, 2020a. Parameterized explainer for graph neural network. In: 34th International Conference on Neural Information Processing Systems, NeurIPS.

Luo, Chenxu, Sun, Lin, Dabiri, Dariush, Yuille, Alan, 2020b. Probabilistic multi-modal trajectory prediction with Lane attention for autonomous vehicles. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS. pp. 2370–2376.

Mandal, Sampurna, Biswas, Swagatam, Balas, Valentina E., Shaw, Rabindra Nath, Ghosh, Ankush, 2020. Motion prediction for autonomous vehicles from lyft dataset using deep learning. In: 2020 IEEE 5th International Conference on Computing Communication and Automation, ICCCA. pp. 768–773.

Mo, Xiaoyu, Xing, Yang, Lv, Chen, 2021. Graph and recurrent neural network-based vehicle trajectory prediction for highway driving. In: 2021 IEEE International Intelligent Transportation Systems Conference, ITSC. pp. 1934–1939.

Pope, Phillip E., Kolouri, Soheil, Rostami, Mohammad, Martin, Charles E., Hoffmann, Heiko, 2019. Explainability methods for graph convolutional neural networks. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 10764–10773.

Rao, Jiahua, Zheng, Shuangjia, Yang, Yuedong, 2022. Quantitative evaluation of explainable graph neural networks for molecular property prediction. Patterns 100628.

Rudin, Cynthia, 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1, 206–215.

Salzmann, Tim, Ivanovic, Boris, Chakravarty, Punarjay, Pavone, Marco, 2020. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In: European Conference on Computer Vision, ECCV.

Schlichtkrull, Michael Sejr, Cao, Nicola De, Titov, Ivan, 2021. Interpreting graph neural networks for NLP with differentiable edge masking. In: International Conference on Learning Representations, ICLR.

Schnake, Thomas, Eberle, Oliver, Lederer, Jonas, Nakajima, Shinichi, Schütt, Kristof T., Müller, Klaus-Robert, Montavon, Grégoire, 2022. Higher-order explanations of graph neural networks via relevant walks. IEEE Trans. Pattern Anal. Mach. Intell. 44 (11), 7581–7596.

Schwarzenberg, Robert, Hübner, Marc, Harbecke, David, Alt, Christoph, Hennig, Leonhard, 2019. Layerwise relevance visualization in convolutional text graph classifiers. In: 13th Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13). pp. 58–62.

Sundararajan, Mukund, Taly, Ankur, Yan, Qiqi, 2017. Axiomatic attribution for deep networks. In: 34th International Conference on Machine Learning.

Tang, Luqi, Yan, Fuwu, Zou, Bin, Li, Wenbo, Lv, Chen, Wang, Kewei, 2023. Trajectory prediction for autonomous driving based on multiscale spatial-temporal graph. IET Intell. Transp. Syst. 17 (2), 255–461.

Office of U.S. Senator Ron Wyden, 2022. Algorithmic accountability act of 2022. 117th Congress 2D Session, https://www.congress.gov/bill/117th-congress/house-bill/6580/text.

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz, Polosukhin, Illia, 2017. Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30.

Veliković, Petar, Cucurull, Guillem, Casanova, Arantxa, Romero, Adriana, Liò, Pietro, Bengio, Yoshua, 2018. Graph attention networks. In: International Conference on Learning Representations, ICLR.

Vu, Minh N., Thai, My T., 2020. PGM-explainer: Probabilistic graphical model explanations for graph neural networks. In: 34th Conference on Neural Information Processing Systems, NeurIPS.

Yang, Yaming, Guan, Ziyu, Li, Jianxin, Zhao, Wei, Cui, Jiangtao, Wang, Quan, 2021. Interpretable and efficient heterogeneous graph convolutional network. IEEE Trans. Knowl. Data Eng. 35 (2), 1637–1650.

Ying, Rex, Bourgeois, Dylan, You, Jiaxuan, Zitnik, Marinka, Leskovec, Jure, 2019. GNNExplainer: Generating explanations for graph neural networks. In: 33rd Conference on Neural Information Processing Systems, NeurIPS.

Yuan, Hao, Tang, Jiliang, Hu, Xia, Ji, Shuiwang, 2020. XGNN: Towards model-level explanations of graph neural networks. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

Yuan, Hao, Yu, Haiyang, Gui, Shurui, Ji, Shuiwang, 2022. Explainability in graph neural networks: A taxonomic survey. IEEE Trans. Pattern Anal. Mach. Intell. 1–19.

Zeng, Wenyuan, Liang, Ming, Liao, Renjie, Urtasun, Raquel, 2021. LaneRCNN: Distributed representations for graph-centric motion forecasting. In: IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS. pp. 532–539.

Zhan, Wei, Sun, Liting, Wang, Di, Shi, Haojie, Clausse, Aubrey, Naumann, Maximilian, Kümmerle, Julius, Königshof, Hendrik, Stiller, Christoph, de La Fortelle, Arnaud, Tomizuka, Masayoshi, 2019. Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps. arXiv:1910.03088.

Zhang, Kunpeng, Feng, Xiaoliang, Wu, Lan, He, Zhengbing, 2022a. Trajectory prediction for autonomous driving using spatial-temporal graph attention transformer. IEEE Trans. Intell. Transp. Syst. 23 (11), 22343–22353.

Zhang, Qingzhao, Hu, Shengtuo, Sun, Jiachen, Chen, Qi Alfred, Mao, Z. Morley, 2022b. On adversarial robustness of trajectory prediction for autonomous vehicles. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR. pp. 15138–15147.

Zhang, Kunpeng, Li, Li, 2022. Explainable multimodal trajectory prediction using attention models. Transp. Res. C 143, 103829.

Zhang, Kunpeng, Zhao, Liang, Dong, Chengxiang, Wu, Lan, Zheng, Liang, 2023. AI-TP: Attention-based interaction-aware trajectory prediction for autonomous driving. IEEE Trans. Intell. Veh. 8 (1), 73–83.

Zhao, Hang, Gao, Jiyang, Lan, Tian, Sun, Chen, Sapp, Benjamin, Varadarajan, Balakrishnan, Shen, Yue, Shen, Yi, Chai, Yuning, Schmid, Cordelia, Li, Congcong, Anguelov, Dragomir, 2020. TNT: Target-driven trajectory prediction. In: 4th Conference on Robot Learning, CoRL.

Zhao, Cong, Song, Andi, Du, Yuchuan, Yang, Biao, 2022. TrajGAT: A map-embedded graph attention network for real-time vehicle trajectory imputation of roadside perception. Transp. Res. C 142, 103787.

Zhdanov, Maksim, Steinmann, Saskia, Hoffmann, Nico, 2022. Investigating brain connectivity with graph neural networks and gnnexplainer. In: 26th International Conference on Pattern Recognition, ICPR. pp. 5155–5161.

Zhou, Yutao, Wu, Huayi, Cheng, Hongquan, Qi, Kunlun, Hu, Kai, Kang, Chaogui, Zheng, Jie, 2021. Social graph convolutional LSTM for pedestrian trajectory prediction. IET Intell. Transp. Syst. 15 (3), 396–405.