

# MTP-GO: Graph-Based Probabilistic Multi-Agent Trajectory Prediction With Neural ODEs

Theodor Westny, Joel Oskarsson, Björn Olofsson and Erik Frisk

The self-archived postprint version of this journal article is available at Linköping University Institutional Repository (DiVA):

<https://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-203164>

N.B.: When citing this work, cite the original publication.

Westny, T., Oskarsson, J., Olofsson, B., Frisk, E., (2023), MTP-GO: Graph-Based Probabilistic Multi-Agent Trajectory Prediction With Neural ODEs, *IEEE Transactions on Intelligent Vehicles*, 8(9), 4223-4236. <https://doi.org/10.1109/TIV.2023.3282308>

Original publication available at:

<https://doi.org/10.1109/TIV.2023.3282308>

<https://www.ieee.org/>

©2023 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

# MTP-GO: Graph-Based Probabilistic Multi-Agent Trajectory Prediction with Neural ODEs

Theodor Westny\* *Student Member, IEEE*, Joel Oskarsson†, Björn Olofsson‡\*, and Erik Frisk\*

**Abstract**—Enabling resilient autonomous motion planning requires robust predictions of surrounding road users’ future behavior. In response to this need and the associated challenges, we introduce our model titled MTP-GO. The model encodes the scene using temporal graph neural networks to produce the inputs to an underlying motion model. The motion model is implemented using neural ordinary differential equations where the state-transition functions are learned with the rest of the model. Multimodal probabilistic predictions are obtained by combining the concept of mixture density networks and Kalman filtering. The results illustrate the predictive capabilities of the proposed model across various data sets, outperforming several state-of-the-art methods on a number of metrics.

**Index Terms**—Trajectory prediction, Neural ODEs, Graph Neural Networks.

## I. INTRODUCTION

**A**UTONOMOUS vehicles are rapidly becoming a real possibility—reaching degrees of maturity that have allowed for testing and deployment on selected public roads [1]. Future progress toward the realization of fully self-driving vehicles still requires human-level social compliance, heavily dependent on the ability to accurately forecast the behavior of surrounding road users. In light of the interconnected nature of traffic participants, in which the actions of one agent can significantly influence the decisions of others, the development of behavior prediction methods is crucial for achieving resilient autonomous motion planning [2]–[4].

As new high-quality data sets continue to emerge and many vehicles already possess significant computing power resulting from vision-based system requirements, the potential for adopting data-driven behavior prediction is increasing. The application of Graph Neural Networks (GNNs) for the considered problem has emerged as a promising approach, partly because of their strong relational inductive bias that facilitates reasoning about relationships within the problem domain [5]. Furthermore, their capacity for multi-agent forecasting is a natural consequence of representing road users as nodes in the graph, enabling simultaneous predictions for multiple targets.

Despite their adaptability and performance in various tasks, deep networks often lack interpretability compared to traditional state-estimation techniques. Recent methods have presented encouraging results by having a deep network compute the

This research was supported by the Strategic Research Area at Linköping-Lund in Information Technology (ELLIIT), the Swedish Research Council via the project *Handling Uncertainty in Machine Learning Systems* (contract number: 2020-04122), and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

\*Department of Electrical Engineering, Linköping University, Sweden.

†Department of Computer and Information Science, Linköping University, Sweden.

‡Department of Automatic Control, Lund University, Sweden.

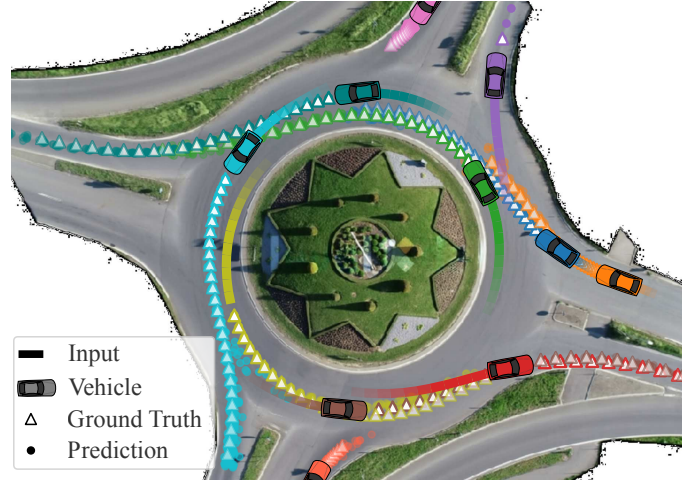


Fig. 1. Example predictions by the proposed model, MTP-GO. Samples drawn from the learned distributions are here used to represent prediction uncertainty. The samples are closely aligned with the ground truth trajectories, illustrating the accuracy and confidence of the model. The example also displays the multimodal capabilities of the method. By studying the samples of the cyan-colored vehicle, two distinct predicted maneuvers can be seen. The image background is part of the *round* data set [12].

inputs to an underlying motion model [6]–[8], convincingly improving interpretability through physically feasible predictions. These approaches are sensible, motivated by the comprehensive literature on motion modeling [9], [10]. However, different road users (e.g., pedestrians, bicycles, and cars) exhibit distinct dynamics, potentially necessitating a collection of customized models. A possible solution is to employ a more general class of methods, such as Neural Ordinary Differential Equations (neural ODEs) [11], to learn inherent differential constraints.

Our main contribution is the proposed MTP-GO<sup>1</sup> model with key properties:

- 1) **Sustained relational awareness:** The model employs a spatio-temporal architecture using specially designed graph-gated recurrent cells that preserve salient inter-agent interactions throughout the complete prediction process.
- 2) **Dynamic versatility:** The model is tailored to adapt to the perpetually dynamic environment while ensuring consistent forecasting for all agents present in the scene at prediction time, irrespective of possible historical information scarcity.
- 3) **Naturalistic predictions:** To compute physically feasible trajectories, the model employs adaptable neural ODEs to learn dynamic motion constraints from data, effectively capturing the inherently smooth nature of physical motion.

<sup>1</sup>Multi-agent Trajectory Prediction by Graph-enhanced neural ODEs

- 4) **Probabilistic forecasting:** To capture the inherent uncertainty and multifaceted nature of traffic, the model combines a mixture density output with an Extended Kalman Filter (EKF) to compute multimodal probabilistic predictions.

The MTP-GO model is compared to several state-of-the-art approaches and evaluated on naturalistic data from various traffic scenarios. Implementations are made publicly available<sup>2</sup>.

## II. RELATED WORK

### A. Traffic Behavior Prediction

The topic of behavior prediction has attracted significant research interest over the past decade. Comprehensive overviews can be found in surveys on the subject [13]–[15]. Broadly, the area can be categorized into two primary streams, focusing on either *intention* or *motion* prediction. Furthermore, a third category may also be recognized—predicting *social patterns*, such as the driving style, attentiveness, or cooperativeness of human drivers [16]. A majority of the methods use sequential input data, like historical agent positions, and often employ Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks [13], for their analyses.

The objective of intention prediction methods is to infer high-level decisions defined by the underlying traffic scene. This includes predicting intention at intersections [17], or lane-change probability in highways [18]. Agent trajectories are labeled according to user-defined maneuvers, and models are trained using a supervised learning approach. While intention prediction is a classification problem, the motion prediction task is regressive by nature, and distance-based measures are typically employed as learning objectives. However, the two prediction problems are not disconnected. This is illustrated in [19], where the predicted intention is used to predict the future trajectory. Because of the sequential nature of the motion prediction problem, several methods base their models on the encoder–decoder framework [20], [21].

Numerous early studies in the field focused on single-agent motion prediction, taking into account only the historical observations of the individual target. The concept of *social pooling*, initially proposed for pedestrian trajectory prediction, was introduced in [22]—among the first works to demonstrate the effectiveness of interaction-aware (IA) modeling. The approach encodes interactions between neighboring agents using pooling tensors. Building on the properties of Convolutional Neural Networks (CNNs), the concept was extended in [23], where the pooling layer was encoded using a CNN to learn the spatial dependencies. The model then uses an RNN-based decoder to generate vehicle trajectories in highway settings. In [24], a more modular approach was proposed, wherein pairwise interactions between agents were learned using multi-head attention mechanisms. Inspired by advancements in sequence prediction in other domains using *Transformers* [25], some researchers have explored their applicability to motion prediction. In [26], a Transformers-based architecture was proposed for the task of pedestrian trajectory prediction. The idea was extended in [27], [28], where IA-mechanisms and

multimodal outputs were achieved using multiple Transformers. Considering the challenges in formulating a supervised learning objective, other studies have investigated the use of inverse reinforcement learning to learn motion prediction tasks [29].

Multimodality and probabilistic predictions are essential to capture the inherent uncertainty in traffic situations. Several methods account for this, most commonly using Mixture Density Networks (MDNs) [23], [24], [30]. Other approaches include generative modeling, such as Conditional Variational Autoencoder (CVAE) [7], [8] and Generative Adversarial Network (GAN) [31].

A potential issue with black-box models is that they might output physically infeasible trajectories. Given the considerable knowledge of motion modeling, it makes intuitive sense to use such knowledge also within data-driven models. In response, some papers have recently included motion constraints within the prediction framework [6]–[8]. Instead of employing pre-defined motion constraints that might be cumbersome to derive or be limited to a single type of road user, the MTP-GO model integrates neural ODEs to learn these constraints directly from the underlying data.

### B. Graph Neural Networks in Motion Prediction

GNNs is a family of deep learning models for graph-structured data [32]. Given a graph structure and a set of associated features, GNNs can be used to learn representations of nodes, edges, or the entire graph [33]. These representations can then be utilized in different prediction tasks. GNNs have demonstrated success in areas such as molecule generation [34], traffic flow prediction [35], and physics simulation [36].

When graph-structured data are collected over time, the resulting samples become time series with associated graphs. Temporal GNNs incorporate additional mechanisms to handle the time dimension. These models can integrate RNNs [35], CNNs [37], or attention mechanisms [8] to model temporal patterns. While most temporal GNNs work with fixed and known graph structures [35], [38], recent studies have also explored learning the graph itself [39].

GNNs can be applied to trajectory prediction problems by letting edges in the graph represent interactions between entities or agents. The LG-ODE model [40] uses an encoder–decoder architecture to predict trajectories of interacting physical objects. A GNN is used to encode historical observations, and a neural ODE decoder predicts the future trajectories. However, this model addresses general physical systems rather than focusing specifically on the traffic setting.

In [41], the use of different GNNs for traffic participant interactions in motion prediction was investigated. Although an early study on the topic, the research showed promising results for IA-modeling. GRIP++ [42] is a graph-structured recurrent model tailored for vehicle trajectory prediction. The scene is encoded to a latent representation using GNN layers [43], which is then passed into an RNN-based encoder–decoder network for trajectory prediction. In [44], SCALE-net was proposed in order to handle any number of interacting agents. Contrasting GRIP++, the node feature updates are encoded using an attention mechanism induced by graph edge features [45]. In [46], node-wise interactions were proposed to be learned

<sup>2</sup><https://github.com/westny/mtp-go>

exclusively using a graph-attention mechanism [47]. The graph encoding is then passed to an LSTM-based decoder for vehicle trajectory prediction. Trajectron++ [7] is a GNN-based method that performs trajectory prediction using a generative model combining an RNN with hard-coded kinematic constraints. Similar to our method, it encodes a traffic situation as a sequence of graphs. STG-DAT [8] is a similarly structured model to that of Trajectron++, and both are considered to be closely related to our model since they utilize temporal GNNs to encode interactions and differentially-constrained motion models to compute the output. Importantly, all above mentioned related works that combine GNNs with some recurrent module only consider the graph during the encoding stages. Motivated by the importance of interaction-aware features, the proposed MTP-GO model instead maintains the graph throughout the entire prediction process, ensuring the retention of key interactions for the full extent of the prediction.

Some research has explored incorporating semantic information in the graph [48], [49]. In [49], the graph is constructed based on semantic goals [30], as opposed to being solely based on the agents themselves. Although this method is not explicitly tailored for trajectory prediction, it provides a more general representation of the scene, which could potentially enhance generalization by transferability. However, the method requires extensive knowledge of map-based information, such as lane lines and signaling signs, which might not always be available, nor does it incorporate agent motion constraints.

### III. PROBLEM DEFINITION

The trajectory prediction problem is formulated as estimating the probability distribution of the future positions  $\mathbf{x}_{t+1}^\nu, \dots, \mathbf{x}_{t+t_f}^\nu$  over the prediction horizon  $t_f$  of all agents  $\nu \in \mathcal{V}_t$  currently in the scene. The model infers the conditional distribution

$$p\left(\left\{\left(\mathbf{x}_{t+1}^\nu, \dots, \mathbf{x}_{t+t_f}^\nu\right)\right\}_{\nu \in \mathcal{V}_t} \middle| \mathcal{H}\right) \quad (1)$$

given the history  $\mathcal{H}$ . The history consists of observations  $\mathbf{f}_i^\nu \in \mathbb{R}^{d_f}$ , such as previous planar positions and velocities from time  $t - t_h$  to  $t$ . Given our focus on the method's future applicability within a predictive controller, the research emphasizes architectural aspects and usability in autonomous motion planning. As a result, we primarily explore lightweight information typically available from vehicle onboard sensors. Based on the outlined problem, this research aims to develop a prediction model that integrates several key properties:

- 1) The model is required to be interaction-aware, learning interactions from historical observations but also preserving them throughout the entire prediction process. The model should therefore predict the future state of the current traffic *scene*, rather than the future states of individual agents independently.
- 2) As urban traffic is characterized by dynamic environments in which various road users enter and exit the vicinity of the ego vehicle, the model must accommodate this variability and consistently predict future trajectories for *all* agents present at the time of prediction. This means

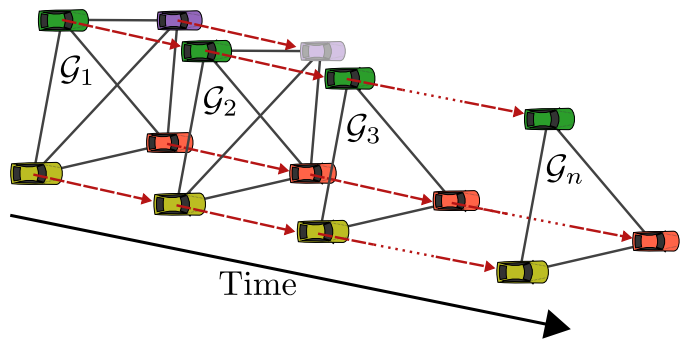


Fig. 2. Example of a sequence of graphs describing a traffic situation starting with  $|\mathcal{V}_1| = 4$  agents. After time 2, the faded agent leaves the traffic situation.

that the model might have varying amounts of data from different agents when making the prediction.

- 3) Vehicle trajectories typically reside on a smooth manifold; thus, the model should compute predictions that are smooth and dynamically feasible. This will be achieved by integrating a dynamic motion model, which is not fixed to a specific type but must be adaptable to different types of road users, e.g., pedestrians, bicycles, or cars.
- 4) Since predictions of intention and motion will be inherently uncertain and multimodal, the model must represent the prediction uncertainty.

### IV. GRAPH-BASED TRAFFIC MODELING

A traffic situation over  $n$  time steps is modeled as a sequence  $\mathcal{G}_1, \dots, \mathcal{G}_n$  of graphs. The node set of the graph  $\mathcal{G}_i$  is  $\mathcal{V}_i$ , corresponding to the agents involved in the traffic situation. The edge set  $\mathcal{E}_i$  is introduced to describe possible edge features. As agents enter or leave the traffic situation over time, the graphs and the cardinality of the node sets change. An example is shown in Fig. 2. Trajectory forecasting is done for all agents  $\mathcal{V}_t$  in the traffic situation at time step  $t$ . The exact observation history of length  $t_h$  can be summarized as

$$\mathcal{H} = \left(\{\mathcal{G}_i\}_{i=t-t_h}^t, \left\{\left\{\mathbf{f}_i^\nu\right\}_{\nu \in \mathcal{V}_i}\right\}_{i=t-t_h}^t\right) \quad (2)$$

Graph construction follows an ego-centric approach, meaning that the input graph is built around a single vehicle; see Fig. 3. This is motivated by the connection to autonomous navigation, where predictions of surrounding vehicles are used for robust ego-vehicle decision-making. Since agents can enter the traffic situation at time steps  $i > t - t_h$ , the feature histories of different nodes can be of different lengths. Regardless, the model should still output a prediction for all nodes in  $\mathcal{V}_t$ , despite possible information scarcity. Motivated by its connection to model predictive control [50], the prediction horizon was set to  $t_f = 5$  s.

#### A. Input Features

The historic observations  $\mathbf{f}_i^\nu$  associated with every node  $\nu$  can be divided into *node* and *context* features. These both refer to various positional information but are separated here since the context features are specific to each data set. Both

TABLE I  
NODE FEATURES

Feature	Description	Unit
$x$	Longitudinal coordinate with respect to $x_0$	m
$y$	Lateral coordinate with respect to $y_0$	m
$v_x$	Instantaneous longitudinal velocity	m/s
$v_y$	Instantaneous lateral velocity	m/s
$a_x$	Instantaneous longitudinal acceleration	m/s <sup>2</sup>
$a_y$	Instantaneous lateral acceleration	m/s <sup>2</sup>
$\psi$	Yaw angle	rad

node and context features are time-varying by nature. In addition, we make use of *static* features, referring to node-specific time-invariant information. Furthermore, edge features are also included in the form of Euclidean distance between the connected nodes.

1) *Node Features*: The time-varying features pertaining to the nodes are summarized in Table I. These include planar positions, velocities, and accelerations. The planar coordinates are given relative to a pre-defined origin  $(x_0, y_0)$ , which is specific for each data set. For highway scenarios, this refers to the graph-centered node's position at the prediction time instant. For urban traffic scenarios, such as roundabouts and intersections, these refer to the roundabout circumcenter and the road junction point of intersection.

2) *Road Context Features*: For highway data, context features add additional information on lateral position with respect to the current lane centerline and the overall road center [51]. The lane position, denoted  $d_l$ , refers to the vehicle's lateral deviation from the current lane center, bounded to the interval  $[-1, 1]$  according to

$$d_l = 2 \frac{y + y_0 - l_{y,l}}{l_w} - 1 \quad (3)$$

where  $l_{y,l}$  is the lateral coordinate of the left lane divider for the current lane and  $l_w$  is the lane width. The road position  $d_r$  is defined similarly, replacing the lane width with the breadth of the road and using the left-most lane divider as a reference. The addition of these features is motivated by their potential to convey information about lane-changing maneuvers [51].

For other scenarios, the context features are simply polar coordinates. These are given with respect to the origin  $(x_0, y_0)$

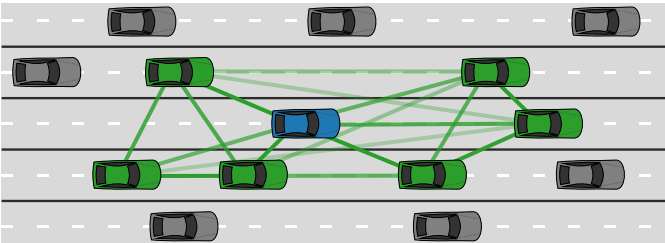


Fig. 3. Graph construction for a highway forecasting problem, centered around a randomly selected vehicle. The graphs are complete with undirected edges and it is the model that learns the importance of each edge. This choice of structure greatly simplifies the implementation.

defined in Section IV-A1 and computed according to:

$$r = \sqrt{(x_0 - x)^2 + (y_0 - y)^2} \quad (4a)$$

$$\theta = \arctan2(y_0 - y, x_0 - x). \quad (4b)$$

3) *Static Features*: In this work, only the agent class (pedestrian, bicycle, car, bus, or truck) is considered for use as a static feature. These are encoded using a *one-hot* scheme and handled separately from the temporal features. Static features are not included by default; a separate study on their effect and usability is presented in Section VI-F.

## V. TRAJECTORY PREDICTION MODEL

The complete MTP-GO model consists of a GNN-based encoder-decoder module that computes the inputs to a motion model for trajectory forecasting. The output is multimodal, consisting of several candidate trajectories  $\hat{x}_{t+1}^j, \dots, \hat{x}_{t+t_f}^j$  for components  $j \in \{1, \dots, M\}$ . In this paper, we set  $M = 8$ , based on an assumption on the number of combined longitudinal and lateral modes in the data. In addition, each candidate is accompanied by a predicted state covariance  $P_{t+1}^j, \dots, P_{t+t_f}^j$  that is estimated using an EKF.

- *Encoder*: The traffic scene history is encoded using a temporal GNN. This module adopts an architecture based on Gated Recurrent Units (GRUs) [21], [52] but replaces learnable weight matrices with graph neural networks.
- *Decoder*: The decoder implements the same underlying structure as the encoder, with an added attention mechanism to learn temporal dependencies. The decoder computes the inputs to the motion model and estimates the process noise used in the EKF.
- *Motion model*: The motion model takes the output of the decoder and predicts the future states of the agents. The dynamics are modeled using neural ODEs [11].

A schematic of the full model is shown in Fig. 4.

### A. Temporal Graph Neural Network Encoder

The encoder takes the history  $\mathcal{H}$  of all agents as input and computes a set of representation vectors useful for predicting future trajectories. Using GRUs, the history is processed sequentially to produce a set of  $d_h$ -dimensional representations  $\{h_i^\nu\}_{i=t-t_h}^t$  for each agent  $\nu \in \mathcal{V}_t$ . A standard GRU cell takes as input at time step  $i$  the features  $f_i^\nu$  and the previous representation  $h_{i-1}^\nu$ . Using these, six new intermediate vectors are computed according to

$$[\kappa_{r,i}^\nu \parallel \kappa_{z,i}^\nu \parallel \kappa_{h,i}^\nu] = W_f f_i^\nu \quad (5a)$$

$$[\xi_{r,i}^\nu \parallel \xi_{z,i}^\nu \parallel \xi_{h,i}^\nu] = W_h h_{i-1}^\nu \quad (5b)$$

where  $W_f \in \mathbb{R}^{3d_h \times d_f}$  and  $W_h \in \mathbb{R}^{3d_h \times d_h}$  are learnable weight matrices, and  $\parallel$  is the concatenation operation. These vectors are then used to compute the representation  $h_i^\nu$  for time step  $i$  as

$$r_i^\nu = \sigma(\kappa_{r,i}^\nu + \xi_{r,i}^\nu + b_r) \quad (6a)$$

$$z_i^\nu = \sigma(\kappa_{z,i}^\nu + \xi_{z,i}^\nu + b_z) \quad (6b)$$

$$\tilde{h}_i^\nu = \phi(\kappa_{h,i}^\nu + r_i^\nu \odot \xi_{h,i}^\nu + b_h) \quad (6c)$$

$$h_i^\nu = (\mathbf{1} - z_i^\nu) \odot \tilde{h}_i^\nu + z_i^\nu \odot h_{i-1}^\nu, \quad (6d)$$

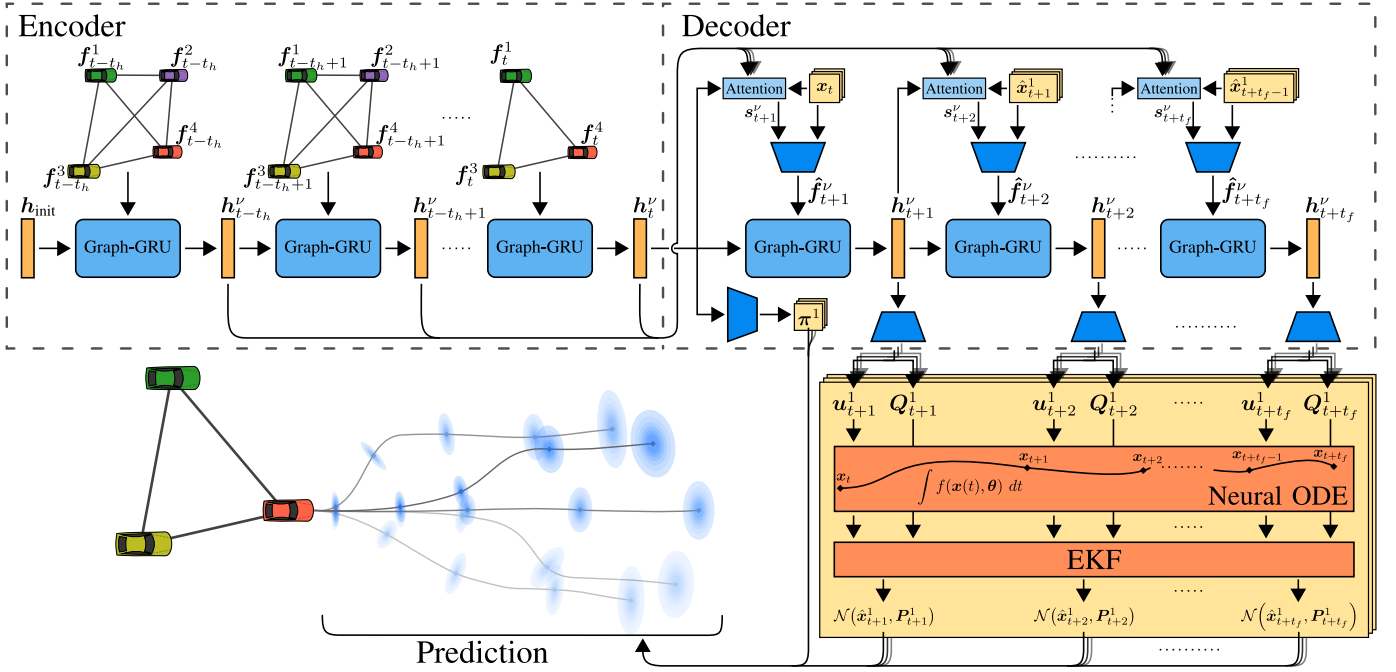


Fig. 4. Schematic illustration of MTP-GO. The figure presents the process of computing predictions for a single agent  $\nu$ . The same process is performed concurrently for all agents in  $\mathcal{V}_t$ .

where the vectors  $\mathbf{b}_r, \mathbf{b}_z, \mathbf{b}_h \in \mathbb{R}_h^d$  are additional bias terms,  $\odot$  is the Hadamard product,  $\sigma$  is the sigmoid function, and  $\phi$  is the hyperbolic tangent. An initial representation  $\mathbf{h}_{\text{init}}$  is used as input for the first encoding step. This vector is learned jointly with other parameters in the model. A standard GRU [52] captures information in the history of agent  $\nu$  but does not incorporate any observations from other agents.

1) *Graph-Gated Recurrent Unit*: In order to accurately predict the future trajectories of each agent, it is important to consider inter-agent interactions. Using the graph formulation in MTP-GO, such interactions can be captured by utilizing an extended GRU cell where the linear mappings in (5) are replaced by GNN components [53], [54]. The GNNs take as input not just the value at the specific node  $\nu$  but also the values of other nodes in the graph. The intermediate representations are computed by two GNNs as

$$[\kappa_{r,i}^\nu \parallel \kappa_{z,i}^\nu \parallel \kappa_{h,i}^\nu] = \text{GNN}_f(\mathbf{f}_i^\nu, \{\mathbf{f}_i^\tau\}_{\tau \neq \nu}) \quad (7a)$$

$$[\xi_{r,i}^\nu \parallel \xi_{z,i}^\nu \parallel \xi_{h,i}^\nu] = \text{GNN}_h(\mathbf{h}_{i-1}^\nu, \{\mathbf{h}_{i-1}^\tau\}_{\tau \neq \nu}). \quad (7b)$$

Note that since the GNN components are built into the GRU cell, the spatial properties of the problem are preserved throughout the encoding and decoding stages. This contrasts [7], [8], [42], [44], [46] where the spatial encoding is done prior to the recurrent operations.

2) *Graph Neural Network Layers*: Each GNN is built by multiple layers, each operating on all nodes concurrently. Here the operation of each layer is described as centered on a node  $\nu$ . Learnable parameters in the GNNs are shared across all nodes but unique to each layer. In this work, multiple types of GNN layers from the literature are adopted:

- *GraphConv* [55] is a straightforward implementation of the Message Passing Neural Network (MPNN) framework [33], which many GNNs are based on. One GraphConv layer computes a new representation  $\mathbf{h}^\nu$  of node  $\nu$  according to

$$\mathbf{h}^\nu = \mathbf{b} + W_1 \mathbf{h}^\nu + \frac{1}{|N(\nu)|} \sum_{\tau \in N(\nu)} e_{\nu,\tau} W_2 \mathbf{h}^\tau \quad (8)$$

where  $W_1, W_2$ , and  $\mathbf{b}$  are learnable parameters,  $e_{\nu,\tau}$  is a weight for the edge  $(\nu, \tau)$ , and  $N(\nu) = \{\tau \mid (\nu, \tau) \in \mathcal{E}\}$  is the neighborhood of node  $\nu$ .

- *Graph Convolutional Network (GCN)* [43] is motivated as a first-order approximation of a learnable spectral graph convolution. The GCN layer update formulation for a node representation is

$$\mathbf{h}^\nu = \mathbf{b} + \sum_{\tau \in \tilde{N}(\nu)} \frac{e_{\nu,\tau}}{\sqrt{d_\nu d_\tau}} W_2 \mathbf{h}^\tau \quad (9)$$

where  $\tilde{N}(\nu) = N(\nu) \cup \{\nu\}$  is the inclusive neighborhood of a node  $\nu$  and  $d_\nu = 1 + \sum_{\tau \in N(\nu)} e_{\nu,\tau}$ .

- *Graph Attention Network (GAT)* [47] layers use an attention mechanism to compute a set of aggregation weights over  $\tilde{N}(\nu)$ . This enables the GNN to focus more on specific neighbors in the graph. The improved GAT [56] version is used, in which attention weights are given by

$$\tilde{\alpha}_{\nu,\tau} = \frac{\exp(\mathbf{a}_{\tilde{\alpha}}^\top \gamma(W_{\tilde{\alpha}}[\mathbf{h}^\nu \parallel \mathbf{h}^\tau] e_{\nu,\tau}))}{\sum_{v \in \tilde{N}(\nu)} \exp(\mathbf{a}_{\tilde{\alpha}}^\top \gamma(W_{\tilde{\alpha}}[\mathbf{h}^\nu \parallel \mathbf{h}^v] e_{\nu,v}))} \quad (10)$$

where  $\mathbf{a}_{\tilde{\alpha}}$  and  $W_{\tilde{\alpha}}$  are learnable parameters, and  $\gamma$  is the leaky ReLU activation function [57]. Note that the edge weight  $e_{\nu,\tau}$  is included as a feature in the computation

of  $\tilde{\alpha}_{\nu,\tau}$ . The new representation is then computed using the attention weights as

$$\mathbf{h}^{\nu} = \mathbf{b} + \sum_{\tau \in \tilde{N}(\nu)} \tilde{\alpha}_{\nu,\tau} W_2 \mathbf{h}^{\tau}. \quad (11)$$

GAT layers can also use multiple *attention heads*, which are independent copies of the attention mechanism described previously. Different heads can pay attention to different aspects, creating multiple separate representation vectors. These vectors are then concatenated or averaged in order to create a new representation.

- *GAT+* is a small extension of GAT where an additional linear transformation is introduced only for the center node

$$\mathbf{h}^{\nu} = \mathbf{b} + W_1 \mathbf{h}^{\nu} + \sum_{\tau \in \tilde{N}(\nu)} \tilde{\alpha}_{\nu,\tau} W_2 \mathbf{h}^{\tau}. \quad (12)$$

This setup introduces additional flexibility in how the representation of the center node is used, something that has shown to be beneficial when evaluating on real data (see Section VI).

3) *Gaussian Kernel Edge Weighting*: Following a commonly used method [35], [38], edge weights are computed using a Gaussian kernel:

$$e_{\nu,\tau} = \exp\left(-\left(\frac{d_{\nu,\tau}}{\sigma_e}\right)^2\right), \quad (13)$$

where  $d_{\nu,\tau}$  is the Euclidean distance between agents  $\nu$  and  $\tau$ . The parameter  $\sigma_e$  controls how much to weigh down edges to agents that are far away. This parameter is learned jointly with the rest of the model.

### B. Decoder with Temporal Attention Mechanism

The decoder also uses a graph-based GRU unit to compute the motion model inputs and process-noise estimates for time steps  $t + 1, \dots, t + t_f$ . As the true sequence of graphs is unknown for these time steps, the decoder always uses the last known graph  $\mathcal{G}_t$  in the GRU. While the  $\mathbf{h}$ -input to these GRU cells functions just as in (7b), the  $\mathbf{f}$ -input is constructed through a temporal attention mechanism [25]. At time  $i > t$ , the attention weights  $\{\alpha_{i,l}^{\nu}\}_{l=t-t_h}^t$  are computed as

$$\mathbf{q}_i^{\nu} = W_{\alpha} [\mathbf{h}_{i-1}^{\nu} \| (W_x \mathbf{x}_{i-1}^{\nu} + \mathbf{b}_x)] + \mathbf{b}_{\alpha} \quad (14a)$$

$$\{\alpha_{i,l}^{\nu}\}_{l=t-t_h}^t = \left\{ \frac{\exp(q_{i,l}^{\nu})}{\sum_{j=1}^{t_h} \exp(q_{i,j}^{\nu})} \right\}_{l=1}^{t_h}, \quad (14b)$$

where  $W_x \in \mathbb{R}^{d_h \times M \cdot d_s}$ ,  $\mathbf{b}_x \in \mathbb{R}^{d_h}$ ,  $W_{\alpha} \in \mathbb{R}^{t_h \times 2d_h}$ , and  $\mathbf{b}_{\alpha} \in \mathbb{R}^{t_h}$  are learnable parameters. For inference, the agent states  $\mathbf{x}_{i-1}^{\nu}$  comes from the prediction of the  $M$  components at the previous time step. During training, teacher forcing [58] is used, where the ground-truth value of  $\mathbf{x}_{i-1}^{\nu}$  is used. The attention weights represent how much attention is paid to the full encoder representation  $\mathbf{o}_i^{\nu} = [\mathbf{h}_{i-t_h}^{\nu} \| \dots \| \mathbf{h}_i^{\nu}]$  at decoder time step  $i$ . The relevant encoded information is then summarized as

$$\mathbf{s}_i^{\nu} = \sum_{l=t-t_h}^t \alpha_{i,l}^{\nu} \mathbf{h}_l^{\nu} \quad (15)$$

and the  $\kappa_{\cdot,i}^{\nu}$ -representations given by

$$\hat{\mathbf{f}}_i^{\nu} = \gamma\left(W_{\hat{f}}[\mathbf{s}_i^{\nu} \| \tilde{\mathbf{x}}_{i-1}^{\nu}] + \mathbf{b}_{\hat{f}}\right) \quad (16a)$$

$$[\kappa_{r,i}^{\nu} \| \kappa_{z,i}^{\nu} \| \kappa_{h,i}^{\nu}] = \text{GNN}_{\hat{f}}\left(\hat{\mathbf{f}}_i^{\nu}, \left\{\hat{\mathbf{f}}_i^{\tau}\right\}_{\tau \neq \nu}\right). \quad (16b)$$

where  $\gamma$  is the leaky ReLU activation function. The new decoder representation  $\mathbf{h}_i^{\nu}$  is finally computed according to (6). This representation is then mapped through additional linear layers to compute the motion model input  $\mathbf{u}$  and parameters defining the process noise matrix  $\mathbf{Q}$  for all  $M$  components. Learnable parameters in the decoder are shared across all time steps.

### C. Motion Model

There exist well-established models used in target tracking [9] and predictive control applications [10] that hold potential for use in trajectory prediction. However, while wheeled vehicles are often well described by nonholonomic constrained models, other road users, like pedestrians, may not share the same constraints. The commonality between agents is that their motion is typically formulated mathematically using ODEs. With the recent proposal of neural ODEs [11], such a flexible formulation could apply here by instead learning the underlying motion model but still enjoying the benefits of smooth trajectories. A neural ODE should learn the parameters  $\boldsymbol{\theta}$  that best describe the state derivative

$$\frac{d\mathbf{x}(t)}{dt} = f(\mathbf{x}(t), \boldsymbol{\theta}), \quad (17)$$

where the states can be retrieved by solving an initial value problem. This part of the network is denoted as the motion model  $f$  and implemented using a fully-connected neural network with ELU activation functions [59]. The function  $f$  accepts two input vectors: the prior state  $\mathbf{x}$  and the current input  $\mathbf{u}$ . Motivated by the degrees of freedom in ground vehicles, the dimensionality of the input  $\mathbf{u}$  is fixed to two. In MTP-GO, both a first and a second-order model are investigated.

1) *First-Order Model*: For the first-order model, there are two model states  $(x, y)$ . Two separate neural ODEs,  $f_1$  and  $f_2$ , are used to describe the state dynamics where each model is associated with its respective input:

$$\begin{aligned} \dot{x} &= f_1(x, y, u_1) \\ \dot{y} &= f_2(x, y, u_2) \end{aligned} \quad (18)$$

2) *Second-Order Model*: The same design principle is employed for higher-order models. The neural ODEs are assigned to model the highest-order state dynamics:

$$\begin{aligned} \dot{x} &= v_x \\ \dot{y} &= v_y \\ \dot{v}_x &= f_1(v_x, v_y, u_1) \\ \dot{v}_y &= f_2(v_x, v_y, u_2) \end{aligned} \quad (19)$$

While it may not share the maneuverability properties of the first-order model, the additional integrations increase the smoothness of the trajectory, which might be useful in some scenarios. For an extended study on motion models in graph-based trajectory prediction, see [60].

#### D. Uncertainty Propagation

For multi-step forecasting, it is reasonable that the estimated uncertainty depends on prior predictions. To provide such an estimate, the time-update step of the EKF is employed. For a given differentiable state-transition function  $f$ , input  $\mathbf{u}_k$ , and process noise  $\mathbf{w}_k$ :

$$\mathbf{x}_{k+1} = f(\mathbf{x}_k, \mathbf{u}_k) + \mathbf{w}_k, \quad (20)$$

the prediction step of the EKF is formulated as:

$$\hat{\mathbf{x}}_{k+1} = f(\hat{\mathbf{x}}_{k|k}, \mathbf{u}_k) \quad (21a)$$

$$\mathbf{P}_{k+1} = \mathbf{F}_k \mathbf{P}_{k|k} \mathbf{F}_k^T + \mathbf{G}_k \mathbf{Q}_{k|k} \mathbf{G}_k^T \quad (21b)$$

where

$$\mathbf{F}_k = \left. \frac{\partial f}{\partial \mathbf{x}} \right|_{\hat{\mathbf{x}}_{k|k}, \mathbf{u}_k}. \quad (22)$$

In (21),  $\hat{\mathbf{x}}$  and  $\mathbf{P}$  refer to the state estimate and state covariance estimate, respectively. The matrix  $\mathbf{Q}_k$  denotes the process noise covariance and should, in combination with  $\mathbf{G}_k$ , describe uncertainties in the state-transition function  $f$ , e.g., because of noisy inputs or modeling errors. Here the state-transition function  $f$  is taken as the current motion model.

To compute the estimated state covariance matrix  $\mathbf{P}_k$ , the decoder is tasked with computing the covariance  $\mathbf{Q}_k$  for every time step  $k$ . For all motion models  $f$ , the process noise  $\mathbf{w}$  is assumed to be a consequence of the predicted inputs and therefore enters into the two highest-order states. Here,  $\mathbf{w}$  is assumed to be zero-mean with covariance matrix

$$\mathbf{Q} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \quad (23)$$

where  $-1 \leq \rho \leq 1$  and  $\sigma_1 > 0, \sigma_2 > 0$ . To get the correct signs of the correlation coefficient and standard deviations, they are passed through Softsign and Softplus activations [58].

How to construct  $\mathbf{G}_k$  depends on the assumptions about the noise  $\mathbf{w}_k$  and how it enters into  $f$ . If it is assumed to be non-additive (cf. (20)), such that  $\mathbf{x}_{k+1} = f(\mathbf{x}_k, \mathbf{u}_k, \mathbf{w}_k)$ , then  $\mathbf{G}_k$  is defined by the Jacobian:

$$\mathbf{G}_k = \left. \frac{\partial f}{\partial \mathbf{w}} \right|_{\hat{\mathbf{x}}_{k|k}, \mathbf{u}_k}. \quad (24)$$

While this is often a design choice, such a formulation is necessary for a completely general neural ODE motion model, where, e.g., each input enters into every predicted state. By the presented modeling approach and assuming the noise is additive,  $\mathbf{G}_k$  can be designed as a matrix of constants. In the simplest case with two state variables, then  $\mathbf{G}_k = T_s \cdot \mathbf{I}_2$ , where  $T_s$  is the sample time. For higher-order state-space models,  $\mathbf{G}_k$  can be generalized as:

$$\mathbf{G}_k = T_s \begin{pmatrix} 0 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (25)$$

An interesting consequence is that this formulation simplifies generating  $\mathbf{P}$  when the number of states  $> 2$ . In several previous works, only the bivariate case is considered. With the proposed approach, the method can be generalized to any number of states as long as  $\mathbf{u}_k$  is 2-dimensional and  $\mathbf{Q}_k$  is modeled explicitly.

#### E. Multimodal Probabilistic Output

Inspired by MDNs [61], [62], the model computes multimodal predictions by learning the parameters of a Gaussian Mixture Model (GMM) for each future time step. Each output vector  $\mathbf{y}_k$  of the model contains mixing coefficients  $\boldsymbol{\pi}^j$ , along with the predicted mean  $\hat{\mathbf{x}}_k^j$  of the states  $\mathbf{x}_k$  and estimated state covariance  $\mathbf{P}_k^j$  for all mixtures  $j \in \{1, \dots, M\}$ :

$$\mathbf{y}_k = \left( \boldsymbol{\pi}^j, \{ \hat{\mathbf{x}}_k^j, \mathbf{P}_k^j \}_{j=1}^M \right). \quad (26)$$

Note that the mixing coefficients  $\boldsymbol{\pi}^j$ , used to represent the weight of the component  $j$ , are constant over the prediction horizon  $t_f$ . For notational convenience, the predictions will be indexed from  $k = 1, \dots, t_f$ . The learning objective is then formulated as the Negative Log-Likelihood (NLL) loss

$$\mathcal{L}_{\text{NLL}} = \sum_{k=1}^{t_f} -\log \left( \sum_j \boldsymbol{\pi}^j \mathcal{N}(\mathbf{x}_k | \hat{\mathbf{x}}_k^j, \mathbf{P}_k^j) \right). \quad (27)$$

1) *Winner Takes All*: MDNs are notoriously difficult to train, and several attempts employ special training mechanisms such as learning parts of the distribution in sequence [63]. Although this might lead to increased stability in the learning process, many MDNs still suffer from mode collapse. The Evolving Winner Takes All (EWTA) loss, proposed in [63], is adopted for the prediction task to address these challenges:

$$\mathcal{L}_{\text{EWTA}}(K) = \sum_{k=1}^{t_f} \sum_{j=1}^M c_j \ell(\hat{\mathbf{x}}_k^j, \mathbf{x}_k) \quad (28a)$$

$$c_j = \delta(j \in B), \quad (28b)$$

where  $\delta(\cdot)$  is the Kronecker delta and  $B$  is the set of component indices pertaining to the current maximum number of winners  $K$  (epoch dependent)

$$B = \underset{\substack{A' \subset A \\ |A'|=K}}{\text{argmin}} \sum_{k=1}^{t_f} \sum_{i \in A'} \ell(\hat{\mathbf{x}}_k^i, \mathbf{x}_k), \quad (29)$$

where  $A = \{1, \dots, M\}$ . This is applied as a loss function during the first training epochs. For fast initial convergence,  $\ell(\cdot)$  was chosen to be the *Huber* loss function [64]—used for winner selection and training.

## VI. EVALUATION & RESULTS

An evaluation of the capabilities of the proposed MTP-GO model was conducted using multiple investigations. Based on the proposals in Section V-A2, a study on different GNN layers and their usability for the considered task is discussed in Section VI-E. Second, an investigation of static features and their usability in the prediction context is presented in Section VI-F. Third, a comparison of the proposed model against related approaches across several data sets and scenarios is presented in Section VI-G. Finally, the results of an ablation study are presented in Section VI-H.

### A. Data Sets

Three different data sets: *highD* [65], *roundD* [12], and *inD* [66] were used for training and testing. The data sets contain recorded trajectories from different locations in Germany, including various highways, roundabouts, and intersections. The data contain several hours of naturalistic driving data recorded at 25 Hz. The input and target data are down-sampled by a factor of 5, effectively setting the sampling time to  $T_s = 0.2$  s. The maximum length of the observation window, i.e., the length of the model input, is set to 3 s, motivated by prior work [67]. Because of the inherent maneuver imbalance, the *highD* data were balanced prior to training using data re-sampling techniques. This was done by oversampling lane-change instances and undersampling lane-keeping instances [51]. The pre-processed *highD*, *roundD*, and *inD* data sets consist of 100404, 29248, and 7820 samples (graph sequences), respectively. We allocate 80% of the total samples for training, 10% for validation, and 10% for testing.

### B. Training and Implementation Details

All implementations were done in PyTorch [68] and for GNN components, PyTorch Geometric [69] was used. Jacobian computations were made efficient by the `functorch` package [70]. The optimizer *Adam* [71] was used with a batch size of 128. Learning rate and hidden dimensionality were tuned using grid search independently for each experiment.

1) *Scheduling of Training Objective*: The model was trained using two loss functions, EWTA and NLL. If  $\mathcal{T}$  is the total number of training epochs, the EWTA is used during an initial warm-up period of  $\mathcal{T}_{\text{warm}} = \mathcal{T}/4$  epochs. Additionally, for the first  $\mathcal{T}_{\text{EWTA}} = \mathcal{T}/8$  epochs the EWTA loss is used exclusively. The exact loss calculation for each epoch  $n$  is described in Algorithm 1.

ALGORITHM 1: SCHEDULING OF TRAINING OBJECTIVE

---

```

1: if  $n < \mathcal{T}_{\text{EWTA}}$  then
2:    $K = \lceil M \cdot (\mathcal{T}_{\text{EWTA}} - n) / \mathcal{T}_{\text{EWTA}} \rceil$  //number of winners
3:    $\mathcal{L} = \mathcal{L}_{\text{EWTA}}(K)$ 
4: else if  $\mathcal{T}_{\text{EWTA}} \leq n < \mathcal{T}_{\text{warm}}$  then
5:    $\beta = (\mathcal{T}_{\text{warm}} - n) / (\mathcal{T}_{\text{warm}} - \mathcal{T}_{\text{EWTA}})$ 
6:    $\mathcal{L} = \beta \cdot \mathcal{L}_{\text{EWTA}}(K = 1) + (1 - \beta) \cdot \mathcal{L}_{\text{NLL}}$ 
7: else
8:    $\mathcal{L} = \mathcal{L}_{\text{NLL}}$ 
9: end if

```

---

### C. Evaluation Metrics

Typically,  $L^2$ -based metrics are used to evaluate prediction performance. However, it is also important to measure the prediction likelihood, as it is an indicator of the model’s ability to capture the uncertainty in its prediction. The metrics are here presented for a single agent. These values are then averaged over all agents in all traffic situations in the test set. For the non-probabilistic metrics, it is important to consider which of the predicted components should be used. Since MTP-GO is punished against mode-collapse during training, it is counter-intuitive to take the average over all components. To that end,

$L^2$ -based metrics are provided with regard to  $\hat{\mathbf{x}}_k$  from the GMM component  $j^*$  with the predicted largest weight

$$j^* = \operatorname{argmax}_j \pi^j \quad (30)$$

Here, the reduced state-vector  $\mathbf{x} = [x, y]$  is used and correspondingly for  $\hat{\mathbf{x}}$ .

- *Average Displacement Error (ADE)*: The average  $L^2$ -norm over the complete prediction horizon is

$$\text{ADE} = \frac{1}{t_f} \sum_{k=1}^{t_f} \|\hat{\mathbf{x}}_k - \mathbf{x}_k\|_2 \quad (31)$$

- *Final Displacement Error (FDE)*: The  $L^2$ -norm of the final predicted position reflects the model’s accuracy in forecasting distant future events:

$$\text{FDE} = \|\hat{\mathbf{x}}_{t_f} - \mathbf{x}_{t_f}\|_2 \quad (32)$$

- *Miss Rate (MR)*: The ratio of cases where the predicted final position is not within 2 m (from [14]) of the ground truth. This indicates prediction consistency.
- *Average Path Displacement Error (APDE)*: The average minimum  $L^2$ -norm between the predicted positions and ground truth is used to estimate the path error. This is used to determine predicted maneuver accuracy:

$$\text{APDE} = \frac{1}{t_f} \sum_{k=1}^{t_f} \|\hat{\mathbf{x}}_k - \mathbf{x}_{k^*}\|_2 \quad (33)$$

$$k^* = \operatorname{argmin}_i \|\hat{\mathbf{x}}_k - \mathbf{x}_i\|_2$$

- *Average Negative Log-Likelihood (ANLL)*: This metric provides an estimate of how well the predicted distribution matches the observed data:

$$\text{ANLL} = \frac{1}{t_f} \sum_{k=1}^{t_f} -\log \left( \sum_j \pi^j \mathcal{N}(\mathbf{x}_k | \hat{\mathbf{x}}_k^j, \mathbf{P}_k^j) \right) \quad (34)$$

It is also useful in determining the correctness of maneuver-based predictions if the method is multimodal.

- *Final Negative Log-Likelihood (FNLL)*: The NLL equivalent of FDE is

$$\text{FNLL} = -\log \left( \sum_j \pi^j \mathcal{N}(\mathbf{x}_{t_f} | \hat{\mathbf{x}}_{t_f}^j, \mathbf{P}_{t_f}^j) \right) \quad (35)$$

### D. Models Compared

The following models are included in the comparative study:

- ❖ *Constant Acceleration (CA)*: Open-loop model assuming constant acceleration.
- ❖ *Constant Velocity (CV)*: Open-loop model assuming constant velocity.
- ❖ *Sequence to Sequence (Seq2Seq)*: Baseline LSTM-based encoder–decoder model (non-interaction-aware).
- ❖ *Social LSTM (S-LSTM)* [22]: Uses an encoder–decoder network based on LSTM for trajectory prediction. Interactions are encoded using social pooling tensors.
- ❖ *Convolutional Social Pooling (CS-LSTM)* [23]: Similar to S-LSTM, but learns interactions using a CNN.

- ❖ Graph Recurrent Network (GNN-RNN) [46]: Encodes interactions using a graph network and generates trajectories with an RNN-based encoder–decoder.
- ❖ mmTransformer [27]: Transformer-based model for multi-modal trajectory prediction. Interactions are encoded using multiple stacked Transformers.
- ❖ Trajectron++ [7]: GNN-based recurrent model. Performs trajectory prediction by a generative model together with hard-coded kinematic constraints.
- ❖ MTP-GO1: Our method with a first-order neural ODE.
- ❖ MTP-GO2: Our method with a second-order neural ODE.

The selection of related methods (S-LSTM, CS-LSTM, GNN-RNN, mmTransformer, and Trajectron++) was based on their relevance and the availability of the authors’ code to the public. Although the implementation of STG-DAT [8], a method closely related to ours, is not publicly accessible, we consider it comparable to Trajectron++ due to their similarities. In order to achieve a fair comparison, the methods were modified to make use of the same input features (see Sections IV-A1 through IV-A2) as MTP-GO, including edge weights for graph-based methods (GNN-RNN and Trajectron++). It was also observed that both mmTransformer and Trajectron++ were highly sensitive to feature scaling, requiring the standardization of input features to have zero mean and unit variance in order to attain desired performance. Apart from these modifications, the models were preserved as per their original proposals and code. All methods were subject to some hyperparameter tuning, specifically using methods to find good learning rates [72], before being trained until convergence.

### E. Parameterizing Graph Neural Network Layers in MTP-GO

Different types of GNN layers can be used in the graph components of MTP-GO. Additionally, the GAT and GAT+ layers can incorporate different numbers of attention heads. These choices can significantly impact the ability of the model to capture interactions in the traffic scenario. In Table II, MTP-GO2 models with different types of GNN layers are compared empirically on the *highD* and *roundD* data sets. The metrics reported here are averaged over all vehicles in the traffic scene.

The results indicate that the best choice for the GNN layers in MTP-GO is GAT+. Also, the comparatively simple GraphConv layer performs surprisingly well. Both of these layers feature some form of parameterization that handles the center node separately from the neighborhood. This seems particularly useful for the roundabout scenario in *roundD*. In general, differences in performance are small for the *highD* data set, but the choice of GNN layer can be crucial for *roundD*. Using multiple attention heads is slightly beneficial for models with GAT layers but does not make a notable difference for GAT+. All types of layers incorporate edge weights in some way. Using these weights has shown to be important for accurately modeling agent interactions. As a comparative example, a model with GCN layers not using edge weights achieves an ADE of 0.86 m on *highD* and 5.61 m on *roundD*. On the *roundD* data set in particular, the edge weights are highly informative, as there can be agents on the other side of the roundabout that do not impact the prediction significantly. For comparison with

TABLE II  
PERFORMANCE OF MTP-GO2 USING DIFFERENT TYPES OF GNN LAYERS

	ADE	FDE	MR	APDE	ANLL	FNLL
<i>highD</i>						
GraphConv	0.29	0.97	<b>0.06</b>	0.28	-1.67	1.61
GCN	0.29	0.99	0.07	0.28	-1.75	1.60
GAT (1 head)	0.32	1.04	0.07	0.31	-1.56	1.68
GAT (3 head)	0.30	0.96	0.07	0.28	-1.57	1.61
GAT (5 head)	0.28	0.91	<b>0.06</b>	0.27	-1.76	1.42
GAT+ (1 head)	0.29	0.94	<b>0.06</b>	0.27	-1.75	1.43
GAT+ (3 head)	0.28	0.91	<b>0.06</b>	0.27	-1.78	1.44
GAT+ (5 head)	<b>0.27</b>	<b>0.90</b>	<b>0.06</b>	<b>0.26</b>	<b>-1.86</b>	<b>1.40</b>
<i>roundD</i>						
GraphConv	1.03	3.28	0.38	0.62	-0.14	3.96
GCN	1.86	5.68	0.62	1.13	1.15	4.95
GAT (1 head)	1.36	4.12	0.44	0.84	0.57	4.27
GAT (3 heads)	1.37	4.08	0.47	0.84	0.67	4.33
GAT (5 heads)	1.24	3.77	0.43	0.78	0.56	4.28
GAT+ (1 head)	<b>0.97</b>	3.05	0.36	<b>0.60</b>	<b>-0.17</b>	<b>3.86</b>
GAT+ (3 heads)	0.98	3.06	<b>0.35</b>	0.61	-0.06	3.88
GAT+ (5 heads)	<b>0.97</b>	<b>3.02</b>	<b>0.35</b>	0.62	-0.03	3.90

other methods in the following sections, MTP-GO using GAT+ layers with one attention head is used.

### F. Static Features in MTP-GO

MTP-GO makes no assumptions about the underlying motion model. Therefore, it could potentially benefit from additional information about the types of agents present in the scene. This was investigated by concatenating the neural-ODE inputs with a one-hot encoding of the agent classes (see Section IV-A3). To fully illustrate their usability, the *inD* data set was used because of its comprehensive content of diverse road users. The resulting prediction performance is presented in Table III, with the CA and CV models as references. Here, the -S suffix adheres to models that include static features. Note that the models are trained for all agent types concurrently; the results are only separated for the test data.

Overall, the first-order model, MTP-GO1, performs the best. Interestingly, its connection to the CV model, the best of the two reference models, illustrates that additional maneuverability is important in this context. The addition of static features does not offer conclusive results. While it does seem to have a minor positive impact on the second-order model, MTP-GO2, the results suggest the opposite for MTP-GO1. However, the most pronounced effect was instead observed during the training process, where models that included static features had a tendency to overfit toward the training data, becoming overconfident in their predictions. Overall this indicates that the use of static features has a potential effect on prediction performance but requires additional research. In the comparative study (Section VI-G), the static features are not included.

### G. Comparative Study

Since only a handful of the considered methods have multi-agent forecasting capabilities, the metrics are provided with regard to the graph-centered vehicle to provide a fair comparison. The methods S-LSTM, CS-LSTM, and MTP-GO offer the possibility to compute the NLL analytically, which is not the

TABLE III  
IND PERFORMANCE PER ROAD USER

Model	Pedestrians		Bicycles		Cars	
	ADE	FDE	ADE	FDE	ADE	FDE
CA	0.72	2.20	1.87	5.95	2.35	7.61
CV	0.58	1.42	2.06	5.40	2.90	7.42
MTP-GO1	<b>0.38</b>	<b>1.03</b>	<b>0.90</b>	3.38	<b>1.07</b>	3.38
MTP-GO1-S	0.39	1.05	<b>0.90</b>	<b>2.65</b>	1.08	<b>3.32</b>
MTP-GO2	0.42	1.13	1.01	3.00	1.27	3.86
MTP-GO2-S	0.40	1.08	1.00	2.98	1.20	3.69

case for the sampling-based Trajectron++. While it might not reflect the true likelihood, the NLL values of Trajectron++ are computed using a kernel density estimate based on samples drawn from the predictive distribution [7] (marked by italics in Tables IV and V). Furthermore, the non-probabilistic metrics of Trajectron++ are computed by averaging over samples drawn from the most likely component. Similarly, the metrics of mmTransformer are computed based on the predicted trajectory with the largest confidence score [27].

1) *Highway*: The performance on the *highD* data set is presented in Table IV. Given the high mean velocity of the vehicles in the data, the challenge of the task lies in predicting over large distances. Despite this, the learning-based methods are very accurate, with a maximum reported FDE of approximately 1.7 m; less than the length of an average car. Out of all methods considered, MTP-GO shows the best performance across most metrics, except in reported NLL. However, the estimated NLL of Trajectron++ might not be comparable to the analytical NLL values of other models. As a comparison, using the kernel density estimation method from Trajectron++ to samples drawn from the predicted distribution of MTP-GO2, the calculated values are  $-7.16$  and  $-3.35$  for the average and final likelihood. This illustrates a potential limitation of sampling-based estimates as they may not accurately represent the true likelihood of the distribution. Although considered methodologically close to MTP-GO, Trajectron++ does not achieve comparable performance on  $L^2$ -based metrics. While the method was not originally intended for the investigated data set, it went through the same hyperparameter tuning as all others considered and trained approximately 4 times longer, possibly attributed to difficulties linked to the CVAE formulation [7]. Still, Trajectron++ achieves one of the lowest values on APDE, indicating that it can predict the correct path, which is possibly attributed to the underlying motion model. Between MTP-GO1 and MTP-GO2, the former achieves slightly better performance. This is interesting considering that real highway driving is typically smooth, which would motivate the use of additional states in MTP-GO2. Moreover, the CA model performs relatively well compared to learning-based methods, possibly indicating that highway trajectory prediction is a simpler prediction problem.

2) *Roundabout*: In Table V, the performance is presented for the *roundD* data set. Upon initial examination, it is evident that predicting roundabout trajectories poses a greater challenge than the highway counterpart. While the significance of interaction-aware modeling concerning the highway prediction problem was not so pronounced, the results are more indicative in this

TABLE IV  
HIGHD PERFORMANCE

Model	ADE	FDE	MR	APDE	ANLL	FNLL
CA	0.78	2.63	0.55	0.73	—	—
CV	1.49	4.01	0.79	1.89	—	—
Seq2Seq	0.57	1.68	0.29	0.54	—	—
S-LSTM [22]	0.41	1.49	0.22	0.39	-0.61	3.20
CS-LSTM [23]	0.39	1.38	0.19	0.37	-0.66	3.33
GNN-RNN [46]	0.40	1.40	0.17	0.38	—	—
mmTransformer [27]	0.39	1.13	0.15	0.39	—	—
Trajectron++ [7]	0.44	1.62	0.23	0.32	<i>(-1.57)</i>	<i>(1.63)</i>
MTP-GO1	<b>0.30</b>	<b>1.07</b>	<b>0.13</b>	<b>0.30</b>	<b>-1.59</b>	<b>2.02</b>
MTP-GO2	0.35	1.16	0.15	0.34	-1.34	2.18

TABLE V  
ROUND PERFORMANCE

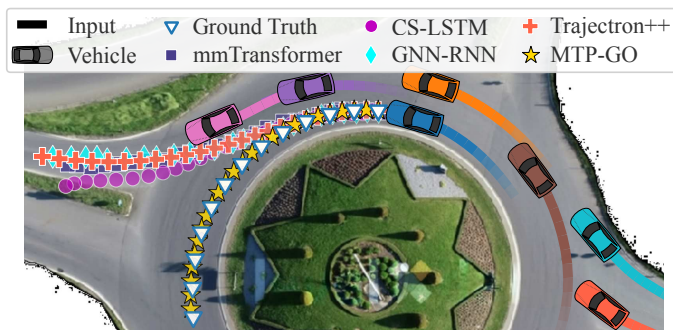
Model	ADE	FDE	MR	APDE	ANLL	FNLL
CA	4.83	16.2	0.95	3.90	—	—
CV	6.49	17.1	0.94	4.34	—	—
Seq2Seq	1.46	3.66	0.59	0.82	—	—
S-LSTM [22]	1.20	3.47	0.56	0.74	1.75	5.12
CS-LSTM [23]	1.19	3.57	0.60	0.69	2.09	5.54
GNN-RNN [46]	1.13	3.11	0.51	0.69	—	—
mmTransformer [27]	1.29	3.50	0.59	0.77	—	—
Trajectron++ [7]	1.09	3.53	0.54	0.59	<i>(-4.25)</i>	<i>(1.50)</i>
MTP-GO1	0.96	<b>2.95</b>	<b>0.46</b>	0.59	0.22	<b>3.38</b>
MTP-GO2	<b>0.92</b>	2.97	0.48	<b>0.57</b>	<b>-0.22</b>	3.85

specific context, clearly favoring graph-modeled interactions. This is further corroborated by the worse performance of the CA, CV, and Seq2Seq models. In Fig. 5, three different scenarios from the roundabout test set are illustrated. Analyzing the prediction quality revealed that several related methods showed competitive performance. However, one key contributor to the improved metrics of MTP-GO is its ability to correctly predict decelerating and yielding maneuvers, an aspect that many other methods struggle with (see Fig. 5b). This is arguably attributed to the interaction-aware properties of MTP-GO and its capacity to preserve the graph. Incorporating differential constraints within the prediction framework was observed to not only stabilize the training process but also provide valuable extrapolation capabilities, enabling the model to generalize beyond observed data. This presents a distinct advantage over conventional neural network models, which can exhibit shortcomings in this regard (see Fig. 5c).

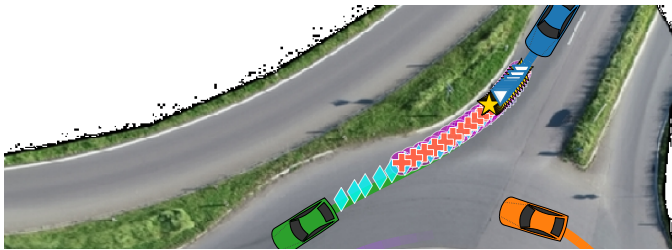
#### H. Ablation Study

The MTP-GO model consists of several components, each with their contribution to the overall performance. In order to get a better understanding of how they affect the model's predictive abilities, an ablation study was conducted. The goal of the ablation study is to dissect the architecture with the objective to determine which mechanisms are most important. The study was conducted for both the *highD* and *roundD* data sets, using MTP-GO2 with single-attention-head GAT+ layers. Three components were considered for the ablation study:

- 1) *Graph Neural Network*: Removing the GNN is comparable to reducing the graph-GRU module to a standard GRU cell. In practice, this is done by removing all edges in the graphs, such that no nodes may share information. This is investigated for both the encoder and decoder.



(a) While many methods achieve good prediction results, MTP-GO is the most accurate overall. In this example scenario, MTP-GO is the only method correct in its maneuver prediction.



(b) Several methods struggle with predicting decelerating maneuvers. Understanding the social queues without knowledge of traffic rules requires substantial interaction-aware capabilities and foresight.



(c) Incorporating differential constraints into prediction models enhances their extrapolation capabilities. By extending the prediction horizon to  $t_f = 7$  s, the models must compute trajectories beyond the observed data (shown by the dashed line). Although this presents a challenge for the majority of models, Trajectron++ and MTP-GO still perform well, which can be ascribed to the differential constraints within these models.

Fig. 5. Example predictions from the roundabout test set using different compared models. Scenarios have been specifically picked to illustrate the properties of the investigated methods. For methods with multimodal capabilities, only the most probable component is shown for clarity. The image background is part of the *round* data set [12].

- 2) *Extended Kalman Filter*: Removing the EKF means that the decoder is tasked with directly computing the state covariance estimate. Although MTP-GO2 has four states, the covariance is only predicted for the first two, i.e., for a bivariate distribution over positions in  $\mathbb{R}^2$ .
- 3) *Neural ODE*: Eliminating the neural ODE removes differential constraints on outputs, requiring direct future state prediction. Note that when the motion model is removed, the EKF cannot be used either (see (21)), such that the complete model reduces to an MDN.

To ensure that the outcomes were not resulting from random variations in model initialization and optimization, each model configuration was trained 10 times using different random seeds. Test performance is presented with an error margin,

based on the 95% confidence interval (CI)

$$CI = \bar{X} \pm t \frac{S}{\sqrt{n}}, \quad (36)$$

where  $\bar{X}$  is the sample mean,  $S$  is the sample standard deviation,  $n = 10$  is the sample size, and  $t = 2.26$  derived from a t-distribution with 9 degrees of freedom [73]. The metrics are computed based on an average of all agents in the scene, similar to the study in Section VI-E. The results are presented in Table VI where a checkmark indicates a component's inclusion in the model whereas a cross indicates exclusion.

In comparison to nominal results, removing the differential constraints impairs the model's effectiveness, although it is more prominent for the highway scenario ( $H_1$  vs.  $H_5$ ). Interestingly, when compared to the removal of other components which have more impact on  $L^2$ -based metrics, the likelihood is seemingly worst affected ( $R_1$  vs.  $R_5$ ). We hypothesize that this is because the model needs to predict a larger variance to accommodate for the added flexibility that comes with the removal of the differential constraints.

Removing the EKF also has a significant effect on the overall performance, seemingly on all metrics (see  $H_4$  and  $R_4$ ). Although it should not directly affect the displacement error, the reason why performance on  $L^2$ -based metrics declines can be attributed to its impact on the learning process. By excluding the EKF, the model is tasked with directly computing the state covariance estimate while attempting to maximize the likelihood. The increased difficulty of the task leads the model to concentrate more on refining covariance estimates rather than minimizing displacement errors. This removal also makes the model more susceptible to random variations, evidenced by the variance in computed likelihood (see  $H_4$  and  $H_7$ ).

This study, along with prior research, emphasizes the significance of modeling interactions for enhancing prediction accuracy. Notably, our work presents a novel contribution by exploring the placement of interaction-aware components within an encoder-decoder framework. Interestingly, in the highway scenario, the placement of GNN components, whether situated in the encoder or decoder, does not appear to impact the results, provided they are present in some capacity ( $H_6$  vs.  $\{H_1, H_2, H_3\}$ ). It is worth pointing out that eliminating the GNN entirely does not substantially diminish performance ( $H_6$ ). This observation is connected to the findings in Section VI-G1, which reveal that although interaction-aware mechanisms enhance prediction performance, the improvement is less prominent in the highway study. In contrast, the roundabout scenario reveals a different outcome—removing the GNN from the decoder leads to a considerable reduction in prediction performance across all metrics ( $R_1$  vs.  $R_3$ ). However, eliminating the GNN from the encoder yields slight, albeit positive, improvements ( $R_2$ ). This is interesting for several reasons. First, if excluding GNN operations from the encoder has minimal or no impact on performance, this translates to a reduced need for parameter optimization and decreased memory requirements. Secondly, this finding is in stark contrast to previous proposals in the context of behavior prediction, where GNN-based models typically incorporate graph operations exclusively during the encoding stages [7], [8], [42], [44], [46].

TABLE VI  
ABLATION STUDY

Data Set	Index	Encoder GNN	Decoder GNN	EKF	ODE	ADE	FDE	MR	APDE	ANLL	FNLL
<i>highD</i>	H <sub>1</sub>	✓	✓	✓	✓	<b>0.28 ± 0.00</b>	<b>0.92 ± 0.01</b>	<b>0.06 ± 0.00</b>	<b>0.27 ± 0.00</b>	-1.74 ± 0.05	1.48 ± 0.05
	H <sub>2</sub>	✗	✓	✓	✓	<b>0.28 ± 0.00</b>	<b>0.92 ± 0.01</b>	<b>0.06 ± 0.00</b>	<b>0.27 ± 0.00</b>	-1.75 ± 0.06	1.47 ± 0.05
	H <sub>3</sub>	✓	✗	✓	✓	<b>0.28 ± 0.00</b>	0.94 ± 0.01	<b>0.06 ± 0.00</b>	<b>0.27 ± 0.00</b>	<b>-1.81 ± 0.03</b>	<b>1.44 ± 0.02</b>
	H <sub>4</sub>	✓	✓	✓	✗	0.29 ± 0.02	0.97 ± 0.03	0.07 ± 0.01	0.28 ± 0.02	-1.76 ± 0.86	1.58 ± 0.28
	H <sub>5</sub>	✓	✓	✓	✗	0.44 ± 0.02	1.15 ± 0.08	0.18 ± 0.03	0.42 ± 0.02	0.63 ± 0.12	2.22 ± 0.05
	H <sub>6</sub>	✗	✗	✗	✓	0.30 ± 0.00	1.03 ± 0.01	0.07 ± 0.00	0.29 ± 0.00	-1.75 ± 0.05	1.61 ± 0.02
	H <sub>7</sub>	✗	✗	✗	✓	0.31 ± 0.02	1.07 ± 0.03	0.08 ± 0.00	0.30 ± 0.01	-1.34 ± 1.12	1.83 ± 0.36
<i>round</i>	R <sub>1</sub>	✓	✓	✓	✓	0.99 ± 0.02	3.10 ± 0.05	<b>0.37 ± 0.02</b>	<b>0.61 ± 0.02</b>	-0.18 ± 0.02	<b>3.78 ± 0.02</b>
	R <sub>2</sub>	✗	✓	✓	✓	<b>0.98 ± 0.01</b>	<b>3.05 ± 0.04</b>	<b>0.37 ± 0.02</b>	<b>0.61 ± 0.02</b>	<b>-0.21 ± 0.02</b>	3.79 ± 0.02
	R <sub>3</sub>	✓	✗	✓	✓	1.07 ± 0.02	3.40 ± 0.06	0.40 ± 0.02	0.63 ± 0.01	-0.10 ± 0.06	3.92 ± 0.03
	R <sub>4</sub>	✓	✓	✗	✓	1.23 ± 0.07	3.83 ± 0.18	0.57 ± 0.03	0.72 ± 0.03	0.20 ± 0.08	4.08 ± 0.07
	R <sub>5</sub>	✓	✓	✗	✗	1.06 ± 0.02	3.05 ± 0.05	0.45 ± 0.03	0.65 ± 0.02	1.60 ± 0.02	3.48 ± 0.01
	R <sub>6</sub>	✗	✗	✓	✓	1.13 ± 0.01	3.64 ± 0.04	0.40 ± 0.02	<b>0.61 ± 0.02</b>	-0.17 ± 0.02	3.95 ± 0.02
	R <sub>7</sub>	✗	✗	✗	✓	1.25 ± 0.01	3.98 ± 0.04	0.51 ± 0.02	0.70 ± 0.01	0.06 ± 0.05	4.03 ± 0.03

## VII. CONCLUSIONS

In this paper, we have presented MTP-GO, a method for probabilistic multi-agent trajectory prediction using an encoder-decoder model based on temporal graph neural networks and neural ordinary differential equations. By incorporating a mixture density network with the time-update step of an extended Kalman filter, the model computes multimodal probabilistic predictions. Key contributions of MTP-GO include its interaction-aware capabilities, attributable to the model’s ability to preserve the graph throughout the prediction process. Additionally, the use of neural ODEs not only enables the model to learn the inherent differential constraints of various road users but also provides it with valuable extrapolation properties that enhance generalization beyond observed data. MTP-GO was evaluated on several naturalistic traffic data sets, outperforming state-of-the-art methods across multiple performance metrics, and showcasing its potential in real-world traffic scenarios.

## ACKNOWLEDGMENT

The authors would like to thank Fredrik Lindsten for helpful discussions and the anonymous reviewers for their valuable suggestions. Computations were enabled by the supercomputing resource Berzelius provided by National Supercomputer Centre at Linköping University and the Knut and Alice Wallenberg foundation.

## REFERENCES

- [1] K. Othman, “Public acceptance and perception of autonomous vehicles: a comprehensive review,” *AI and Ethics*, vol. 1, no. 3, pp. 355–387, 2021.
- [2] T. Brüdigam, J. Zhan, D. Wollherr, and M. Leibold, “Collision avoidance with stochastic model predictive control for systems with a twofold uncertainty structure,” in *IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 432–438.
- [3] I. Batkovic, U. Rosolia, M. Zanon, and P. Falcone, “A robust scenario MPC approach for uncertain multi-modal obstacles,” *IEEE Control Systems Letters*, vol. 5, no. 3, pp. 947–952, 2021.
- [4] J. Zhou, B. Olofsson, and E. Frisk, “Interaction-aware motion planning for autonomous vehicles with multi-modal obstacle uncertainties using model predictive control,” *arXiv preprint arXiv:2212.11819*, 2022.
- [5] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner *et al.*, “Relational inductive biases, deep learning, and graph networks,” *arXiv preprint arXiv:1806.01261*, 2018.
- [6] H. Cui, T. Nguyen, F.-C. Chou, T.-H. Lin, J. Schneider, D. Bradley, and N. Djuric, “Deep kinematic models for kinematically feasible vehicle trajectory predictions,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 10 563–10 569.
- [7] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, “Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data,” in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 683–700.
- [8] J. Li, H. Ma, Z. Zhang, J. Li, and M. Tomizuka, “Spatio-temporal graph dual-attention network for multi-agent prediction and tracking,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 10 556–10 569, 2022.
- [9] X. R. Li and V. P. Jilkov, “Survey of maneuvering target tracking. Part I. Dynamic models,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 39, no. 4, pp. 1333–1364, 2003.
- [10] B. Paden, M. Čáp, S. Z. Yong, D. Yershov, and E. Frazzoli, “A survey of motion planning and control techniques for self-driving urban vehicles,” *IEEE Transactions on Intelligent Vehicles*, vol. 1, no. 1, pp. 33–55, 2016.
- [11] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, “Neural ordinary differential equations,” in *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [12] R. Krajewski, T. Moers, J. Bock, L. Vater, and L. Eckstein, “The round dataset: A drone dataset of road user trajectories at roundabouts in germany,” in *23rd IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2020, pp. 1–6.
- [13] S. Mozaffari, O. Y. Al-Jarrah, M. Dianati, P. Jennings, and A. Mouzakitis, “Deep learning-based vehicle behavior prediction for autonomous driving applications: A review,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 33–47, 2020.
- [14] 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.
- [15] J. Fang, F. Wang, P. Shen, Z. Zheng, J. Xue, and T.-s. Chua, “Behavioral intention prediction in driving scenes: A survey,” *arXiv preprint arXiv:2211.00385*, 2022.
- [16] M. H. Alkinani, W. Z. Khan, and Q. Arshad, “Detecting human driver inattentive and aggressive driving behavior using deep learning: Recent advances, requirements and open challenges,” *IEEE Access*, vol. 8, pp. 105 008–105 030, 2020.
- [17] D. J. Phillips, T. A. Wheeler, and M. J. Kochenderfer, “Generalizable intention prediction of human drivers at intersections,” in *IEEE Intelligent Vehicles Symposium (IV)*, 2017, pp. 1665–1670.
- [18] D. Lee, Y. P. Kwon, S. McMains, and J. K. Hedrick, “Convolution neural network-based lane change intention prediction of surrounding vehicles for ACC,” in *20th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2017.
- [19] L. Xin, P. Wang, C.-Y. Chan, J. Chen, S. E. Li, and B. Cheng, “Intention-aware long horizon trajectory prediction of surrounding vehicles using dual LSTM networks,” in *21st IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 1441–1446.
- [20] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27, 2014.
- [21] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using

- rnn encoder-decoder for statistical machine translation,” *arXiv preprint arXiv:1406.1078*, 2014.
- [22] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, “Social LSTM: Human trajectory prediction in crowded spaces,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 961–971.
- [23] N. Deo and M. M. Trivedi, “Convolutional social pooling for vehicle trajectory prediction,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2018, pp. 1468–1476.
- [24] K. Messaoud, I. Yahiaoui, A. Verroust-Blondet, and F. Nashashibi, “Attention based vehicle trajectory prediction,” *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 1, pp. 175–185, 2021.
- [25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [26] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso, “Transformer networks for trajectory forecasting,” in *25th IEEE international conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 10 335–10 342.
- [27] Y. Liu, J. Zhang, L. Fang, Q. Jiang, and B. Zhou, “Multimodal motion prediction with stacked transformers,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 7577–7586.
- [28] Z. Huang, X. Mo, and C. Lv, “Multi-modal motion prediction with transformer-based neural network for autonomous driving,” in *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2605–2611.
- [29] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, “Deep inverse reinforcement learning for behavior prediction in autonomous driving: Accurate forecasts of vehicle motion,” *IEEE Signal Processing Magazine*, vol. 38, no. 1, pp. 87–96, 2020.
- [30] Y. Hu, W. Zhan, and M. Tomizuka, “Probabilistic prediction of vehicle semantic intention and motion,” in *IEEE Intelligent Vehicles Symposium (IV)*, 2018, pp. 307–313.
- [31] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, “Social GAN: Socially acceptable trajectories with generative adversarial networks,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2255–2264.
- [32] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, “A comprehensive survey on graph neural networks,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, 2021.
- [33] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, “Neural message passing for quantum chemistry,” in *34th International Conference on Machine Learning (ICML)*, 2017, pp. 1263–1272.
- [34] C. Zang and F. Wang, “MoFlow: An invertible flow model for generating molecular graphs,” in *26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, ser. KDD ’20. Association for Computing Machinery, 2020, pp. 617–626. [Online]. Available: <https://doi.org/10.1145/3394486.3403104>
- [35] Y. Li, R. Yu, C. Shahabi, and Y. Liu, “Diffusion convolutional recurrent neural network: Data-driven traffic forecasting,” in *International Conference on Learning Representations (ICLR)*, 2018.
- [36] Y. Rubanova, A. Sanchez-Gonzalez, T. Pfaff, and P. Battaglia, “Constraint-based graph network simulator,” in *39th International Conference on Machine Learning*. PMLR, 2022, pp. 18 844–18 870, ISSN: 2640-3498.
- [37] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, “Graph wavenet for deep spatial-temporal graph modeling,” in *28th International Joint Conference on Artificial Intelligence (IJCAI)*, 2019, p. 1907–1913.
- [38] B. Yu, H. Yin, and Z. Zhu, “Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting,” in *27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018, pp. 3634–3640.
- [39] X. Zhang, M. Zeman, T. Tsiligkaridis, and M. Zitnik, “Graph-guided network for irregularly sampled multivariate time series,” in *International Conference on Learning Representations (ICLR)*, 2022.
- [40] Z. Huang, Y. Sun, and W. Wang, “Learning continuous system dynamics from irregularly-sampled partial observations,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [41] F. Diehl, T. Brunner, M. T. Le, and A. Knoll, “Graph neural networks for modelling traffic participant interaction,” in *IEEE Intelligent Vehicles Symposium (IV)*, 2019, pp. 695–701.
- [42] 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.
- [43] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *International Conference on Learning Representations (ICLR)*, 2017.
- [44] H. Jeon, J. Choi, and D. Kum, “SCALE-Net: Scalable vehicle trajectory prediction network under random number of interacting vehicles via edge-enhanced graph convolutional neural network,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 2095–2102.
- [45] L. Gong and Q. Cheng, “Exploiting edge features for graph neural networks,” in *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019, pp. 9211–9219.
- [46] X. Mo, Y. Xing, and C. Lv, “Graph and recurrent neural network-based vehicle trajectory prediction for highway driving,” in *24th IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2021, pp. 1934–1939.
- [47] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, “Graph attention networks,” in *International Conference on Learning Representations (ICLR)*, 2018.
- [48] M. Zipfl, F. Hertlein, A. Rettinger, S. Thoma, L. Halilaj, J. Luettin, S. Schmid, and C. Henson, “Relation-based motion prediction using traffic scene graphs,” in *25th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2022, pp. 825–831.
- [49] Y. Hu, W. Zhan, and M. Tomizuka, “Scenario-transferable semantic graph reasoning for interaction-aware probabilistic prediction,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 12, pp. 23 212–23 230, 2022.
- [50] V. Fors, B. Olofsson, and E. Frisk, “Resilient branching MPC for multi-vehicle traffic scenarios using adversarial disturbance sequences,” *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 4, pp. 838–848, 2022.
- [51] T. Westny, E. Frisk, and B. Olofsson, “Vehicle behavior prediction and generalization using imbalanced learning techniques,” in *24th IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2021, pp. 2003–2010.
- [52] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, “On the properties of neural machine translation: Encoder–decoder approaches,” in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2014, pp. 103–111.
- [53] X. Zhao, F. Chen, and J.-H. Cho, “Deep learning for predicting dynamic uncertain opinions in network data,” in *IEEE International Conference on Big Data (Big Data)*, 2018, pp. 1150–1155.
- [54] J. Oskarsson, P. Sidén, and F. Lindsten, “Temporal graph neural networks for irregular data,” in *26th International Conference on Artificial Intelligence and Statistics*, 2023.
- [55] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe, “Weisfeiler and leman go neural: Higher-order graph neural networks,” in *AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 4602–4609.
- [56] S. Brody, U. Alon, and E. Yahav, “How attentive are graph attention networks?” in *International Conference on Learning Representations (ICLR)*, 2022.
- [57] A. L. Maas, A. Y. Hannun, A. Y. Ng *et al.*, “Rectifier nonlinearities improve neural network acoustic models,” in *International Conference on Machine Learning (ICML)*, vol. 30, no. 1. Atlanta, Georgia, USA, 2013, p. 3.
- [58] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016.
- [59] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs),” *arXiv e-prints*, p. arXiv:1511.07289, Nov. 2015.
- [60] T. Westny, J. Oskarsson, B. Olofsson, and E. Frisk, “Evaluation of differentially constrained motion models for graph-based trajectory prediction,” in *IEEE Intelligent Vehicles Symposium (IV)*, 2023.
- [61] C. M. Bishop, “Mixture density networks,” Aston University, Tech. Rep., 1994.
- [62] —, *Pattern recognition and machine learning*. Springer, 2006.
- [63] O. Makansi, E. Ilg, O. Cicek, and T. Brox, “Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7144–7153.
- [64] P. J. Huber, “Robust estimation of a location parameter,” in *Breakthroughs in statistics*. Springer, 1992, pp. 492–518.
- [65] R. Krajewski, J. Bock, L. Kloecker, and L. Eckstein, “The highD dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems,” in *21st IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 2118–2125.
- [66] J. Bock, R. Krajewski, T. Moers, S. Runde, L. Vater, and L. Eckstein, “The inD dataset: A drone dataset of naturalistic road user trajectories at german intersections,” in *IEEE Intelligent Vehicles Symposium (IV)*, 2020, pp. 1929–1934.

- [67] S. Yoon and D. Kum, “The multilayer perceptron approach to lateral motion prediction of surrounding vehicles for autonomous vehicles,” in *IEEE Intelligent Vehicles Symposium (IV)*, 2016, pp. 1307–1312.
- [68] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimeshein, L. Antiga *et al.*, “PyTorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [69] M. Fey and J. E. Lenssen, “Fast graph representation learning with PyTorch Geometric,” in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.
- [70] R. Z. Horace He, “functorch: Jax-like composable function transforms for pytorch,” <https://github.com/pytorch/functorch>, 2021, accessed on 23.11.2022.
- [71] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in *International Conference on Learning Representations (ICLR)*, 2015.
- [72] L. N. Smith, “Cyclical learning rates for training neural networks,” in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017, pp. 464–472.
- [73] R. E. Walpole, R. H. Myers, S. L. Myers, and K. Ye, *Probability and statistics for engineers and scientists*, 9th ed. Boston, Massachusetts: Prentice Hall, 2012.



**Theodor Westny** received the B.Sc. degree in Electrical Engineering in 2018 and the M.Sc. degree in Mechatronics 2020, both from Linköping University, Sweden. He is currently pursuing a Ph.D. degree in Electrical Engineering. His research interest includes data-driven behavior prediction, vehicle dynamics modeling, and predictive motion control of autonomous heavy vehicles.



**Joel Oskarsson** received the M.Sc. degree in Computer Science and Engineering in 2020 from Linköping University. He is currently pursuing a Ph.D. degree in Computer Science with a focus on probabilistic machine learning. His research interests include machine learning methods for spatio-temporal and graph-structured data.



**Björn Olofsson** received the M.Sc. degree in Engineering Physics in 2010 and the Ph.D. degree in Automatic Control in 2015, both from Lund University, Sweden. He is currently an Associate Professor at the Department of Automatic Control, Lund University, Sweden, and also affiliated with the Department of Electrical Engineering, Linköping University, Sweden. His research includes motion control for robots and vehicles, optimal control, system identification, and statistical sensor fusion.



**Erik Frisk** was born in Stockholm, Sweden, in 1971. He received the Ph.D. degree in Electrical Engineering in 2001 from Linköping University, Sweden. He is currently a Professor at the Department of Electrical Engineering, Linköping University, Sweden. His main research interests are optimization techniques for autonomous vehicles in complex traffic scenarios as well as model and data-driven fault diagnostics and prognostics.