Towards Generalizable and Interpretable Motion Prediction: A Deep Variational Bayes Approach

Juanwu Lu [1] Wei Zhan [2] Masayoshi Tomizuka [2] Yeping Hu [3]

[1]Purdue University [2]University of California, Berkeley [3]Lawrence Livermore National Laboratory

Summary

This paper proposes the Goal-based Neural Variational Agent (*GNeVA*), an interpretable generative model for motion prediction with robust generalizability to out-of-distribution cases. Experiments on motion prediction datasets validate that the fitted model can be interpretable and generalizable and can achieve comparable performance to state-of-the-art results.

Motivations

a. Modeling the uncertain and multi-modal driver behaviors for motion prediction.
b. Improve limited model generalizability: Performance degradation facing Out-of-Distribution (OOD) data [1].
c. Improve limited model interpretability: Most state-of-the-art methods propose end-to-end black-box prediction models.

Problem Formulation

- **Observation:** Observe surroundings in the previous H time steps.
- **Prediction:** Predict a target agent's future T-step trajectory.
- **Environment Semantics:** The set $\mathcal{S}$ of objects in the surroundings besides traffic participants (e.g., road geometry, traffic regulations).
- **Traffic Participants:** The set $\mathcal{P}$ of individuals or entities interacting in the current traffic (e.g., vehicles, cyclists, and pedestrians). A subset of them $\mathcal{T}$ are targets to predict.
- **Objective:** Find the optimal model $f \in \mathcal{F}$ that parameterizes a probabilistic model that maximizes the likelihood of the target agent's future states

$$\max_{f\in\mathcal{F}} \prod_{n=1}^{|\mathcal{T}|}\prod_{t=1}^{T} p\left(\boldsymbol{x}_{H+t}^{(i)} \mid f(\boldsymbol{x}_{<H+t}^{(j)}, \boldsymbol{s}_{<H+t}^{(k)}); j \in \mathcal{P}, k \in \mathcal{S}\right)$$

- **Target-driven Motion Prediction:** Reduce the problem into two stages: first sample from a continuous spatial distribution over plausible future trajectory endpoints (i.e., *goals*), and then complete the intermediate trajectory [2].

Assumptions and Design Ideas

- **Multimodal Goal Distribution:** Assume the goals follow a mixture of Gaussian.
- **Disentangled Conditional Posteriors for Generalization:** Assume the means and precision matrices follow a Normal-Wishart conjugate prior distribution to improve model generalizability.
  - **Posterior of mean** is conditioned on the environment semantics and all other participants $\boldsymbol{s}$.
  - **Posterior of precision** is only conditioned on other participants $\boldsymbol{x}_{\leq H}$.

References

[1] Juanwu Lu, Wei Zhan, Masayoshi Tomizuka, and Yeping Hu. Generalizability analysis of graph-based trajectory predictor with vectorized representation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13430–13437, 2022.

[2] Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Ben Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, Congcong Li, and Dragomir Anguelov. Tnt: Target-driven trajectory prediction. In Jens Kober, Fabio Ramos, and Claire Tomlin, editors, *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pages 895–904. PMLR, 16–18 Nov 2021.

[3] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B. Choy, Philip H. S. Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[4] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In Leslie Pack Kaelbling, Danica Kragic, and Komei Sugiura, editors, *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 86–99. PMLR, 30 Oct–01 Nov 2020.

[5] Maosheng Ye, Tongyi Cao, and Qifeng Chen. Tpcn: Temporal point cloud networks for motion forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11318–11327, June 2021.

[6] Zhiyu Huang, Xiaoyu Mo, and Chen Lv. Multi-modal motion prediction with transformer-based neural network for autonomous driving. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2605–2611, 2022.

[7] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 541–556, Cham, 2020. Springer International Publishing.

Variational Structure of the Spatial Distribution of Goals

The endpoint position $g \in \mathbb{R}^2$ of a target agent's future trajectory is assumed to follow a Bayesian mixture of Gaussian distributions. As illustrated in Figure 1, we utilizes an unconditional generative process in the following form with learnable prior parameters:

$$\boldsymbol{z} \sim \prod_{n=1}^{N}\prod_{c=1}^{C} \pi_c^{z_{nc}},$$

$$\boldsymbol{g} \mid \boldsymbol{z}, \boldsymbol{\mu}, \boldsymbol{\Lambda} \sim \prod_{n=1}^{N}\prod_{c=1}^{C} \mathcal{N}\left(g_n \mid \boldsymbol{\mu}_c, \boldsymbol{\Lambda}_c^{-1}\right)^{z_{nc}},$$

$$\boldsymbol{\mu}_c, \boldsymbol{\Lambda}_c \sim \mathcal{N}\left(\boldsymbol{\mu}_c \mid \boldsymbol{\eta}_0, (\beta_0\boldsymbol{\Lambda}_c)^{-1}\right)\mathcal{W}\left(\boldsymbol{\Lambda}_c \mid V_0, \nu_0\right).$$

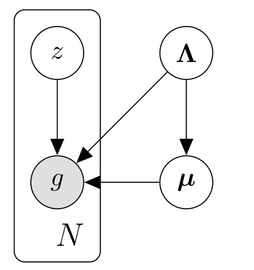

Figure 1. Likelihood Family

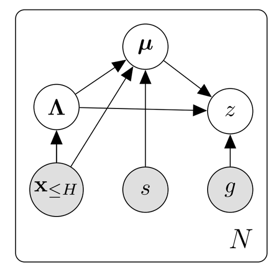

Figure 2. Variational Family

However, the mean vectors and precision matrices should be conditioned on observed history information $\boldsymbol{x}_{\leq H}$ and $\boldsymbol{s}$. Therefore, we learn a set of functions to parameterize the mean-field variational distributions of $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$ and evaluate the distribution of $\boldsymbol{z}$ by

$$\log q(z_{nc}) \approx \mathbb{E}_{q(\boldsymbol{\mu},\boldsymbol{\Lambda})}[\log p(g_n \mid \boldsymbol{\mu}_c, \boldsymbol{\Lambda}_c^{-1}, z_{nc})].$$

Goal-based Neural Variational Agents (GNeVA)

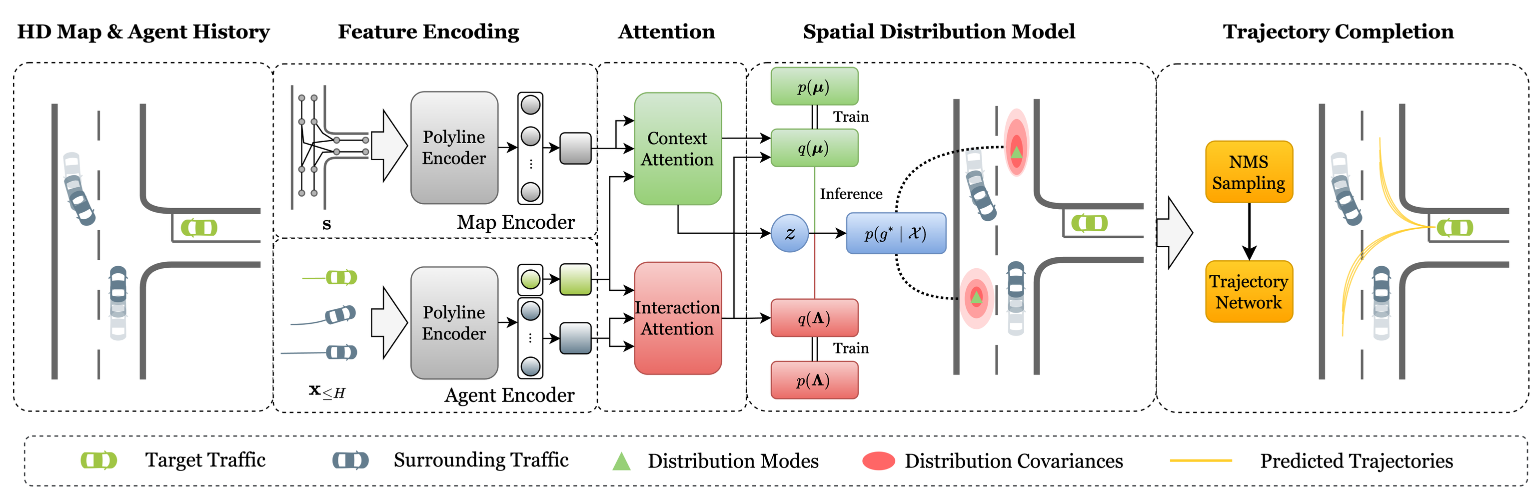

Figure 3. Overview of the GNeVA framework.

- **Feature Encoding:** The traffic scenario is represented as a collection of polylines. We encode map polylines and participants' history trajectories by two separate encoders, resulting in three features: map features $\boldsymbol{m}$, target participant's history feature $\boldsymbol{e}$, and surrounding participants' history feature $\boldsymbol{o}$.
- **Attention Modules** Model global interactions and parameterize the *posterior distributions* of $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$:
  - **Context Attention** uses $\boldsymbol{e}$ as query, concat $[\boldsymbol{m}, \boldsymbol{o}]$ as key and value, and outputs parameters in $q(\boldsymbol{\mu})$.
  - **Interaction Attention** uses $\boldsymbol{e}$ as query, concat $[\boldsymbol{e}, \boldsymbol{o}]$ as key and value, and output parameters in $q(\boldsymbol{\Lambda})$.
- **Proxy z-posterior Network:** An additional module trained to estimate the variational posterior distribution of $\boldsymbol{z}$ using history features:

$$\tilde{p}\left(\boldsymbol{z} \mid \boldsymbol{x}_{\leq H}, \boldsymbol{s}\right) = \text{MLP}(\text{concat}\left[\boldsymbol{x}_{\leq H}, \boldsymbol{s}\right]) \approx q(\boldsymbol{z})$$

- **Sampling and Trajectory Completion:** Sample goals from the following posterior predictive Student's t-distribution using Non-Maximum Suppression

$$p(g^*) \approx \sum_{c=1}^{C} \tilde{p}(\boldsymbol{z})\text{St}_{\nu_c-1}\left(\eta_c, \frac{\beta_c+1}{\beta_c(\nu_c-1)}V_c^{-1}\right).$$

We use a cascade of MLPs for each sampled goal to complete the intermediate trajectories with the goal and the context attention module output as inputs.

Benchmark Results

Table 1. Results on INTERACTION validation set.

	mADE_6	mFDE_6
DESIRE [3]	0.32	0.88
MultiPath [4]	0.30	0.99
TNT [2]	**0.21**	<u>0.67</u>
GNeVA (Ours)	<u>0.25</u>	**0.64**

Table 2. Results on Argoverse validation set.

	mADE_6	mFDE_6	MR_6
TPCN [5]	<u>0.73</u>	1.15	0.11
mmTrans [6]	**0.71**	1.15	0.11
LaneGCN [7]	**0.71**	<u>1.08</u>	-
GNeVA (Ours)	0.78	**1.06**	**0.10**

Generalizability Analysis

Table 3. Model Performance under Cross-scenario Tests

	Train Scenario					
	Intersection		Roundabout		Full Dataset	
Validate Scenario	mADE_6	mFDE_6	mADE_6	mFDE_6	mADE_6	mFDE_6
Intersection	0.56	1.41	0.56	1.39	0.31	0.73
Roundabout	0.61	1.56	0.44	1.08	0.32	0.76

Table 4. Cross Dataset Evaluation Results.

Dataset	Argoverse (validate)			INTERACTION (validate)	
	mADE_6	mFDE_6	MR_6	mADE_6	mFDE_6
Argoverse (train)	0.78	1.06	0.10	0.37	0.91
INTERACTION (train)	0.92	1.34	0.15	0.25	0.64

Visualizations on In-distribution (ID) and OOD Cases

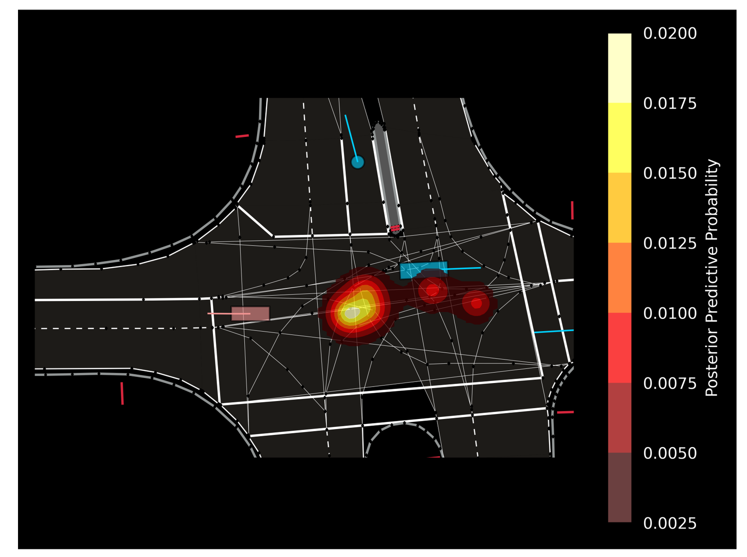

Figure 4. ID case: `USA_Intersection_MA`

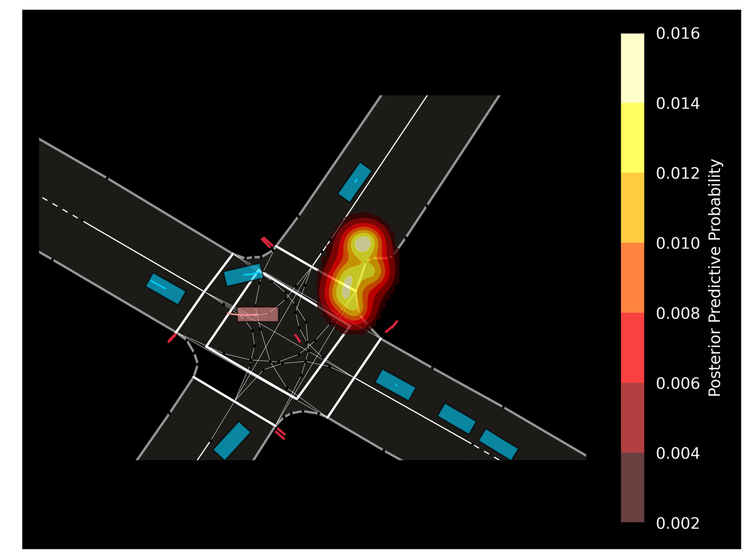

Figure 5. OOD case: `Intersection_CM`

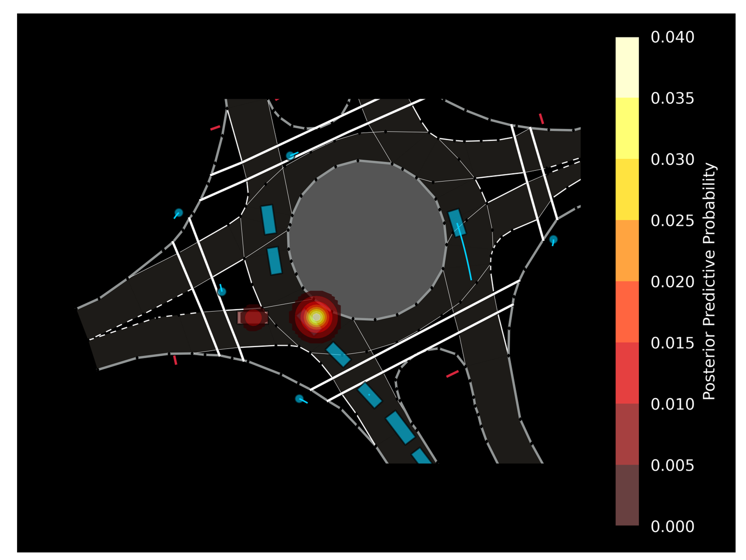

Figure 6. ID case: `USA_Roundabout_SR`

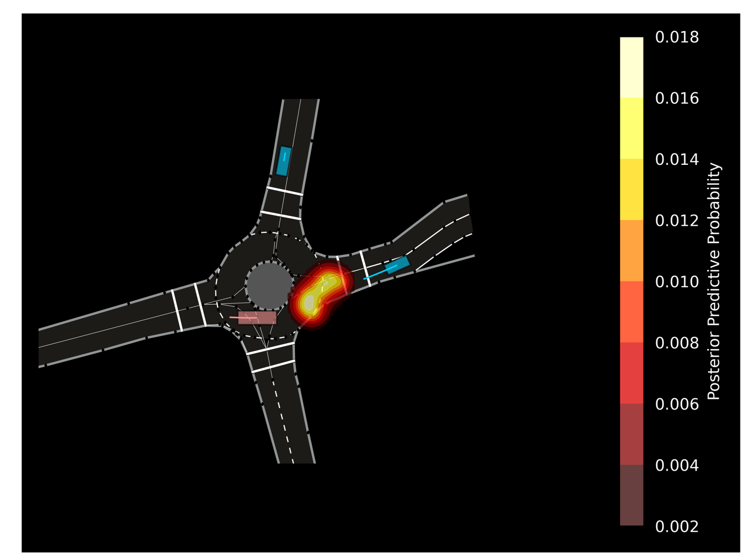

Figure 7. OOD case: `Roundabout_RW`