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

May 2, 2024

Table of Contents

1 Introduction

2 Methodology

3 Results

4 Conclusion

Motion Prediction

Motion prediction in autonomous vehicles (AVs) tries to estimate the future motion states of other traffic participants for a certain period using historical observation about their surroundings.

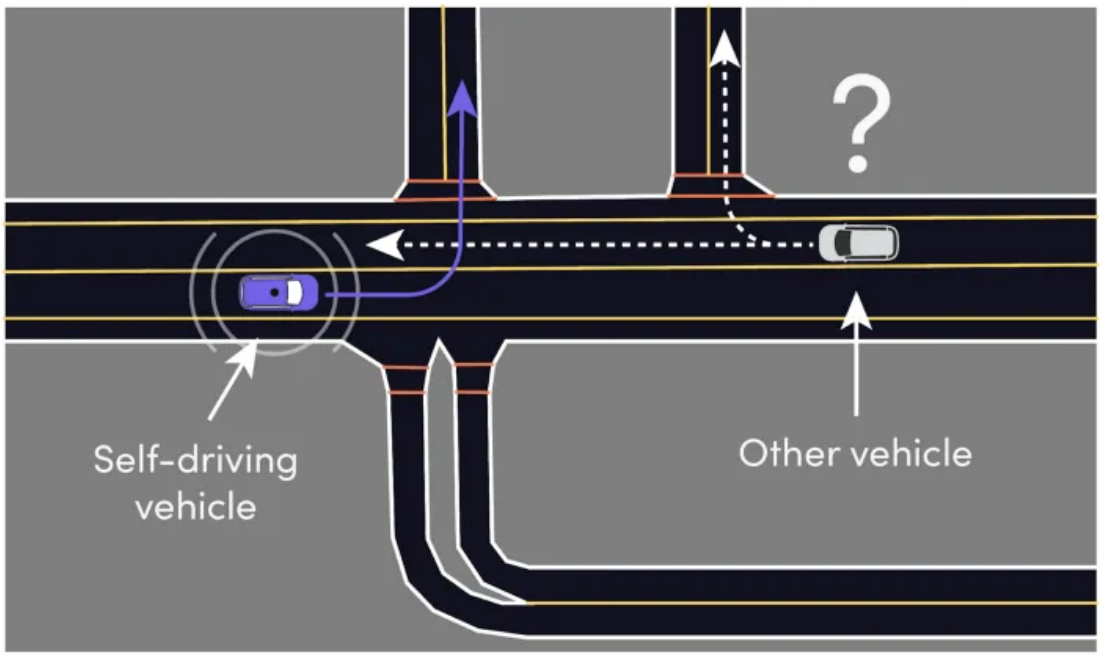


Figure: Illustration of motion prediction. For the AV to perform an unprotected left turn, it needs to know whether the oncoming vehicle will turn right or go straight and interfere with the AV's left turn.
Source: Woven by Toyota.

Motivations

- **Existing Works**
  - Black-box end-to-end models with limited interpretability.
  - The generalizability of the trained model is not fully investigated.
- **Questions**
  - How to model the multimodal trajectory distributions with high-degree uncertainty?
  - How to improve the generalizability of the model?
  - How to incorporate interpretable latent variables in the model?

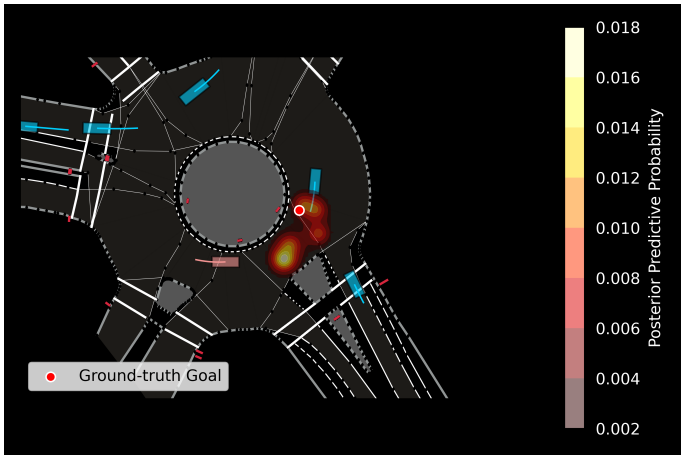


Figure: Illustration of the multimodal distribution of future trajectory endpoints with uncertainty heatmap.

Table of Contents

1 Introduction

2 Methodology

3 Results

4 Conclusion

Ideas

Observation: Cumulative Uncertainty

The destination of a future trajectory accounts for most uncertainty due to cumulative uncertainty over time.

Ideas

Observation: Cumulative Uncertainty

The destination of a future trajectory accounts for most uncertainty due to cumulative uncertainty over time.

Observation: Hierarchical Decision Process

Drivers usually first decide where to go and then adjust their maneuvers to reach that destination.

Ideas

Observation: Cumulative Uncertainty

The destination of a future trajectory accounts for most uncertainty due to cumulative uncertainty over time.

Observation: Hierarchical Decision Process

Drivers usually first decide where to go and then adjust their maneuvers to reach that destination.

Approach: Target-driven Motion Prediction

Formulate the motion prediction problem into two stages: sample plausible destinations and then complete the intermediate trajectories from the current location to these destinations.

Ideas

Objective: Improved Interpretability

Assume the trajectory endpoints follow a mixture of Gaussian distributions, where means and precisions can directly reflect expectations and uncertainty.

Ideas

Objective: Improved Interpretability

Assume the trajectory endpoints follow a mixture of Gaussian distributions, where means and precisions can directly reflect expectations and uncertainty.

Objective: Improved Generalizability

Use conjugate Normal-Wishart prior for the Gaussian parameters and construct a **disentangled conditional posterior**:

Ideas

Objective: Improved Interpretability

Assume the trajectory endpoints follow a mixture of Gaussian distributions, where means and precisions can directly reflect expectations and uncertainty.

Objective: Improved Generalizability

Use conjugate Normal-Wishart prior for the Gaussian parameters and construct a **disentangled conditional posterior**:

- Posterior of mean $\boldsymbol{\mu}$ is conditioned on all other traffic participants' road geometry and history trajectories.

Ideas

Objective: Improved Interpretability

Assume the trajectory endpoints follow a mixture of Gaussian distributions, where means and precisions can directly reflect expectations and uncertainty.

Objective: Improved Generalizability

Use conjugate Normal-Wishart prior for the Gaussian parameters and construct a **disentangled conditional posterior**:

- Posterior of mean $\boldsymbol{\mu}$ is conditioned on all other traffic participants' road geometry and history trajectories.
- Posterior of precision $\boldsymbol{\Lambda}$ is only conditioned on the history trajectories of other traffic participants.

How to model the target/goal distribution?

- **Mixture Model** The goal of a predicted participant $\boldsymbol{g}$ is assumed to follow a Bayesian mixture of Gaussian distributions.

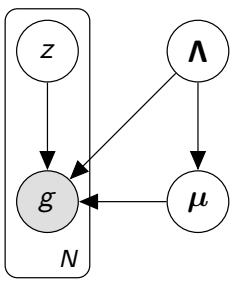


Figure: Likelihood Family

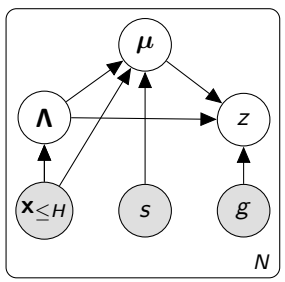


Figure: Variational Family

How to model the target/goal distribution?

- **Mixture Model** The goal of a predicted participant $\boldsymbol{g}$ is assumed to follow a Bayesian mixture of Gaussian distributions.
- **Context Conditioning** The recognition model is conditioned on the environment semantics $\boldsymbol{s}$ and history interaction derived from $\boldsymbol{x}_{\leq H}$.

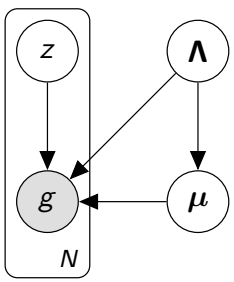


Figure: Likelihood Family

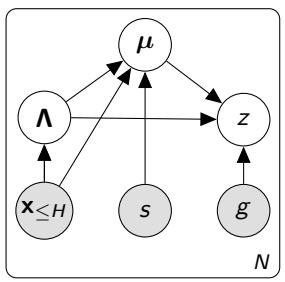


Figure: Variational Family

How to model the target/goal distribution?

- **Mixture Model** The goal of a predicted participant $\boldsymbol{g}$ is assumed to follow a Bayesian mixture of Gaussian distributions.
- **Context Conditioning** The recognition model is conditioned on the environment semantics $\boldsymbol{s}$ and history interaction derived from $\boldsymbol{x}_{\leq H}$.
- **Training** Applying Variational EM Algorithm

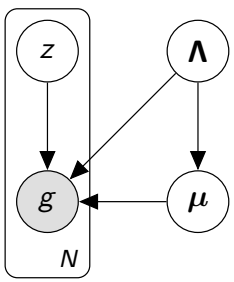


Figure: Likelihood Family

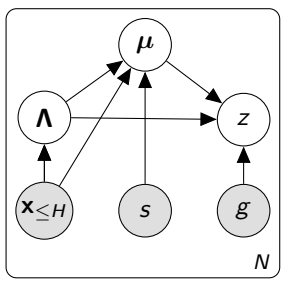


Figure: Variational Family

How to model the target/goal distribution?

- **Mixture Model** The goal of a predicted participant $\boldsymbol{g}$ is assumed to follow a Bayesian mixture of Gaussian distributions.
- **Context Conditioning** The recognition model is conditioned on the environment semantics $\boldsymbol{s}$ and history interaction derived from $\boldsymbol{x}_{\leq H}$.
- **Training** Applying Variational EM Algorithm
  - **E Step**: Evaluate mixture weights $\log q(z_{nk})$ for each batch data n.

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

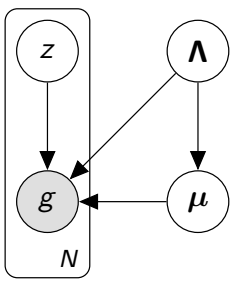


Figure: Likelihood Family

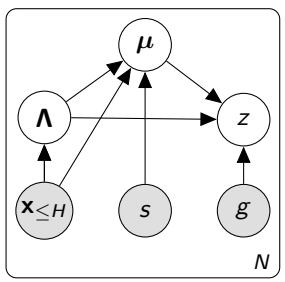


Figure: Variational Family

How to model the target/goal distribution?

- **Mixture Model** The goal of a predicted participant $\boldsymbol{g}$ is assumed to follow a Bayesian mixture of Gaussian distributions.
- **Context Conditioning** The recognition model is conditioned on the environment semantics $\boldsymbol{s}$ and history interaction derived from $\boldsymbol{x}_{\leq H}$.
- **Training** Applying Variational EM Algorithm
  - **E Step**: Evaluate mixture weights $\log q(z_{nk})$ for each batch data n.

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

  - **M Step**: Maximize the Evidence Lower Bound and update network parameters by Stochastic Gradient Descent.

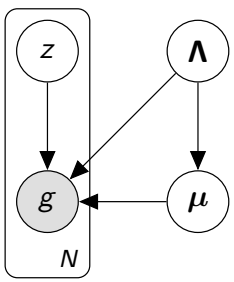


Figure: Likelihood Family

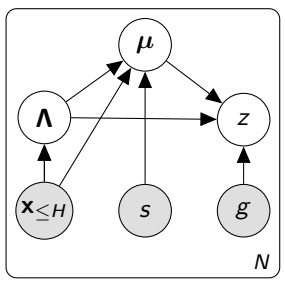


Figure: Variational Family

How to encode the context of the surroundings?

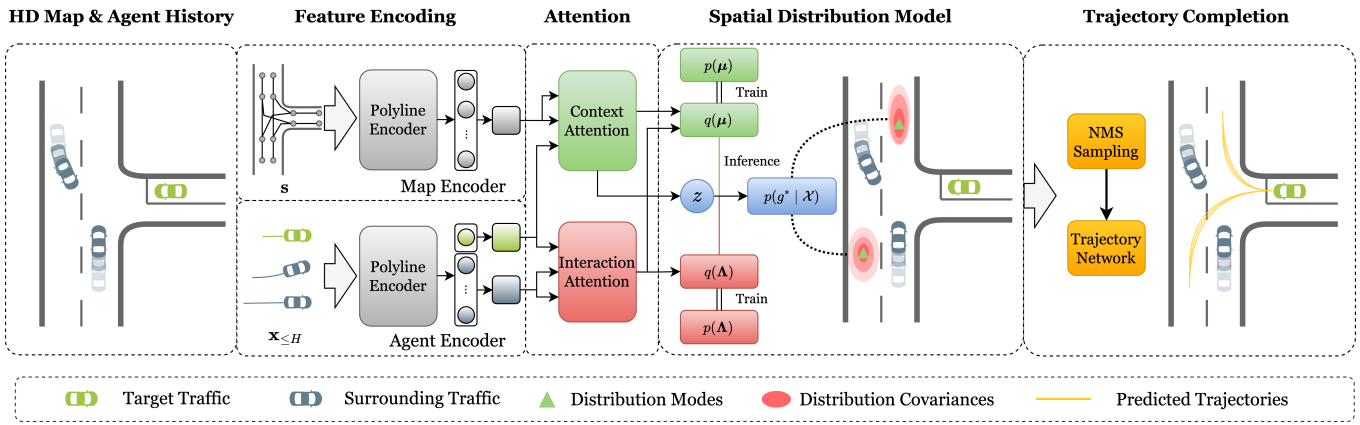


Figure: GNeVA architecture.

- **Input Context Features**
  - Road geometry (road boundaries, lane markings, etc.)
  - History trajectories of other participants.
- **Vectorized Representation**: Represents both of the above as collections of vectors on polylines.
- **Polyline Encoders**: Encode map features ***m***, target participant's history feature ***e***, and other participants' history features ***o***.

How to encode the context of the surroundings?

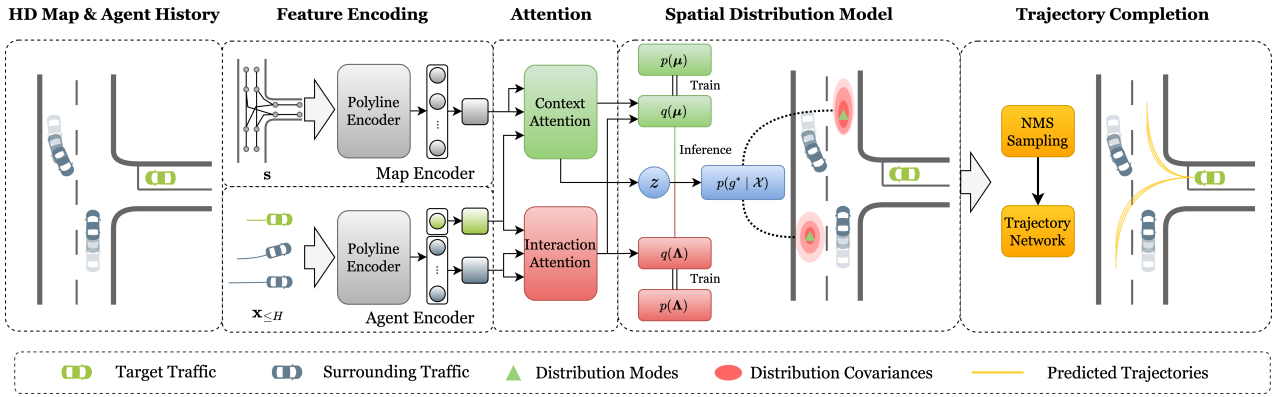


Figure: GNeVA architecture.

- **Attention Modules**

How to encode the context of the surroundings?

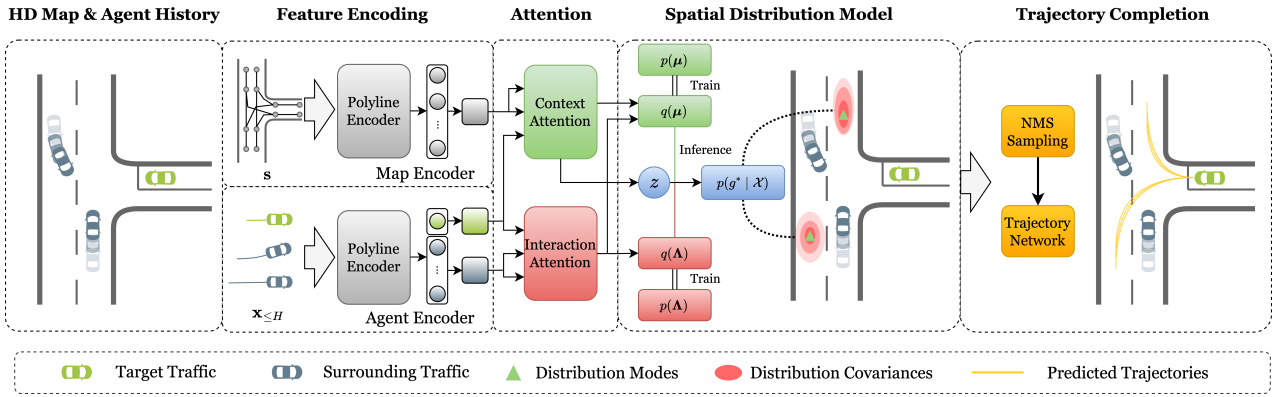


Figure: GNeVA architecture.

- **Attention Modules**
  - Objective: Model global interactions and parameterize the *posterior distributions* of μ and $\mathbf{\Lambda}$.

How to encode the context of the surroundings?

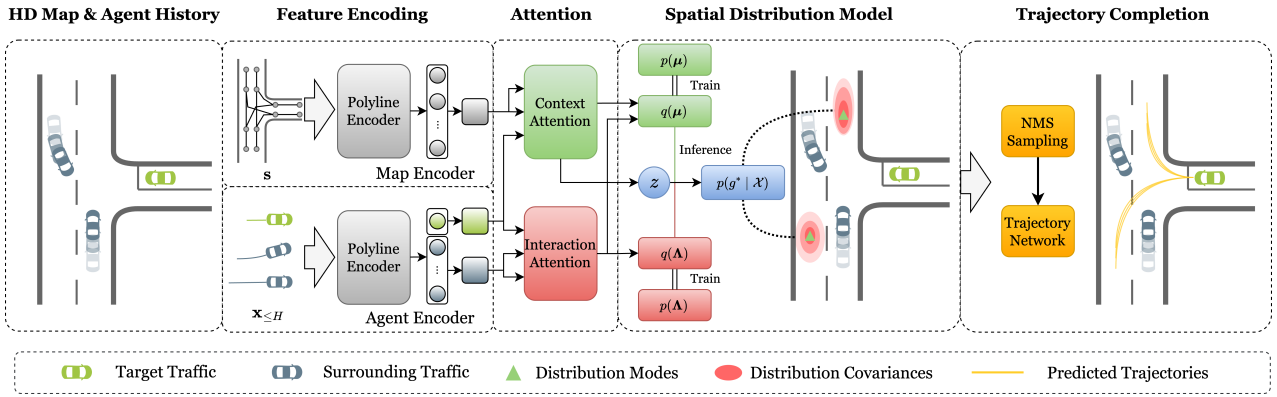


Figure: GNeVA architecture.

- **Attention Modules**
  - Objective: Model global interactions and parameterize the *posterior distributions* of μ and $\mathbf{\Lambda}$.
  - Implementation: A cascade of transformer encoder blocks.

How to encode the context of the surroundings?

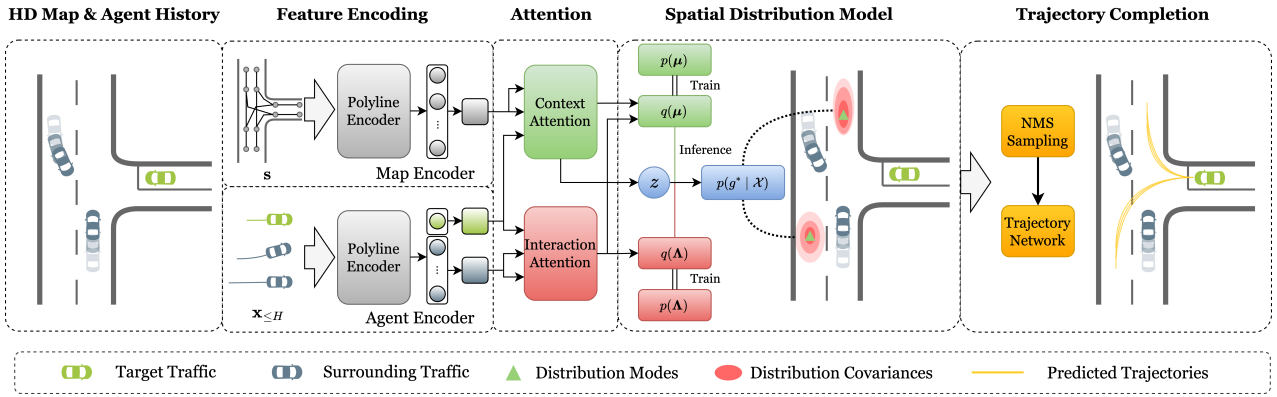


Figure: GNeVA architecture.

- **Attention Modules**
  - Objective: Model global interactions and parameterize the *posterior distributions* of $\boldsymbol{\mu}$ and $\boldsymbol{\Lambda}$.
  - Implementation: A cascade of transformer encoder blocks.
  - Context Attention: uses $\boldsymbol{e}$ as query, concat$[\boldsymbol{m}, \boldsymbol{o}]$ as key and value.

How to encode the context of the surroundings?

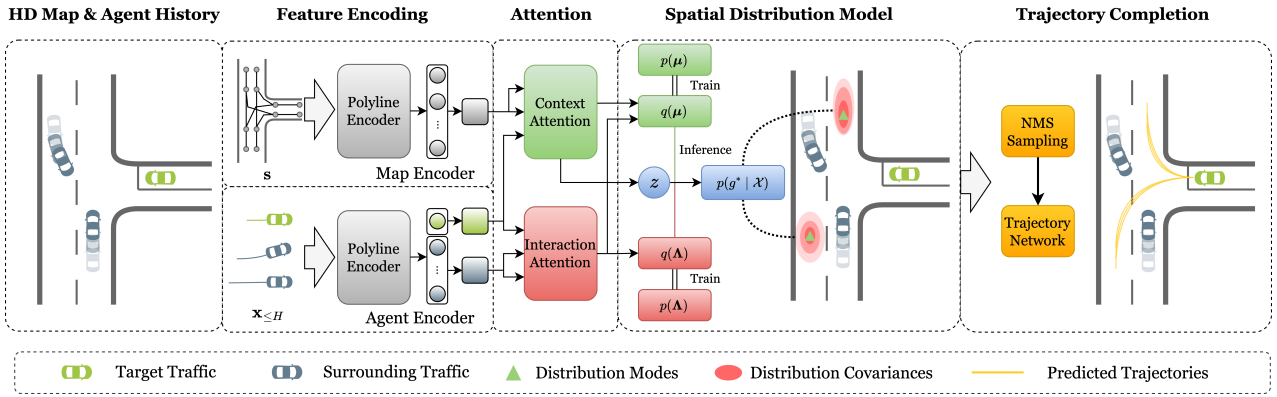


Figure: GNeVA architecture.

- **Attention Modules**
  - Objective: Model global interactions and parameterize the *posterior distributions* of μ and $\mathbf{\Lambda}$.
  - Implementation: A cascade of transformer encoder blocks.
  - Context Attention: uses $\boldsymbol{e}$ as query, concat$[\boldsymbol{m}, \boldsymbol{o}]$ as key and value.
  - Interaction Attention: uses $\boldsymbol{e}$ as query, concat$[\boldsymbol{e}, \boldsymbol{o}]$ as key and value.

How to improve sample efficiency?

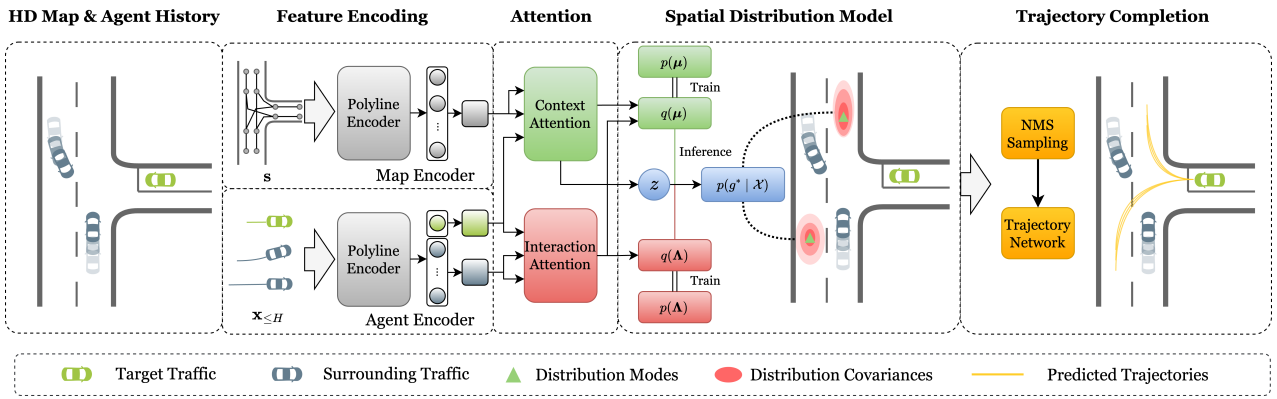


Figure: GNeVA architecture.

- **Objective:** Determine the optimal mixture distributions to sample.
- **Proposal:** An additional module, *proxy z-posterior network*, trained to estimate the variational posterior distribution of $\boldsymbol{z}$:

$$\tilde{p}(\boldsymbol{z} \mid \boldsymbol{x}_{\leq H}, \boldsymbol{s}) \approx q(\boldsymbol{z})$$

- **Training:** Minimizing the cross-entropy loss.

How to sample and generate the full trajectory?

- **Sampling Destination** Apply Non-maximum Suppression on the mixture of multivariate Student distributions.

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

- **Trajectory Completion** 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.

Table of Contents

1 Introduction

2 Methodology

3 Results

4 Conclusion

Benchmarks

- The GNeVA model can achieve performance comparable to existing models.

Table: Results on INTERACTION valid set.

	mADE$_6$	**mFDE$_6$**
DESIRE	0.32	0.88
MultiPath	0.30	0.99
TNT	**0.21**	<u>0.67</u>
GNeVA (Ours)	<u>0.25</u>	**0.64**

Table: Results on Argoverse valid set.

	mADE$_6$	**mFDE$_6$**	**MR$_6$**
TPCN	<u>0.73</u>	1.15	0.11
mmTrans	**0.71**	1.15	0.11
LaneGCN	**0.71**	<u>1.08</u>	-
GNeVA (Ours)	0.78	**1.06**	**0.10**

Model Generalizability

- The GNeVA model can maintain its performance when applied to an unseen scenario.

Table: 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

- The GNeVA model can maintain its performance on a different dataset.

Table: 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: In-distribution (ID) and OOD cases

Figure: ID case from `CHN_Merging_ZS0`

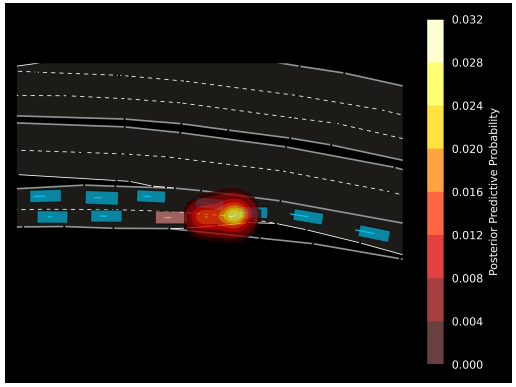


Figure: OOD case from `Merging_TR0`

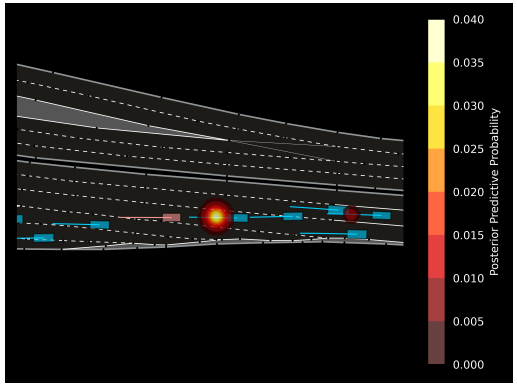
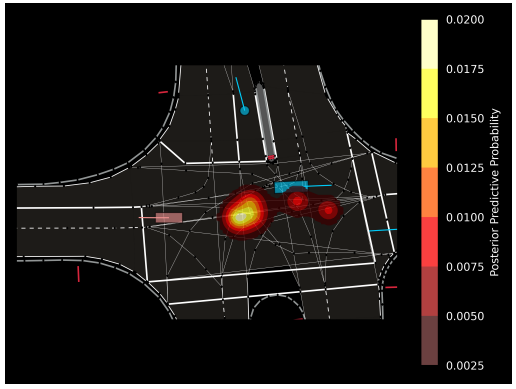


Figure: ID case from `USA_Intersection_MA`



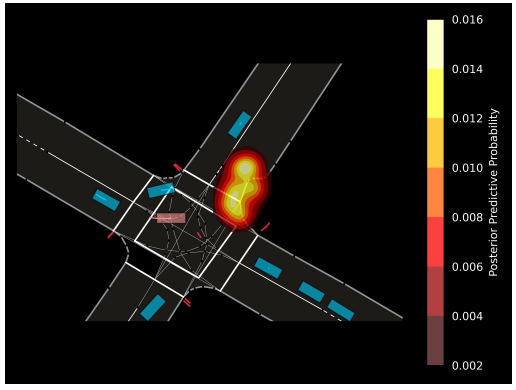


Figure: OOD case from `Intersection_CM`

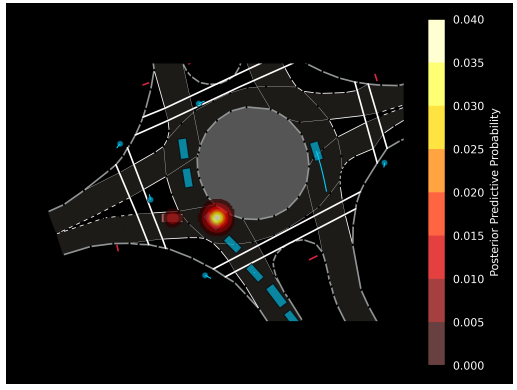


Figure: ID case from `USA_Roundabout_SR`

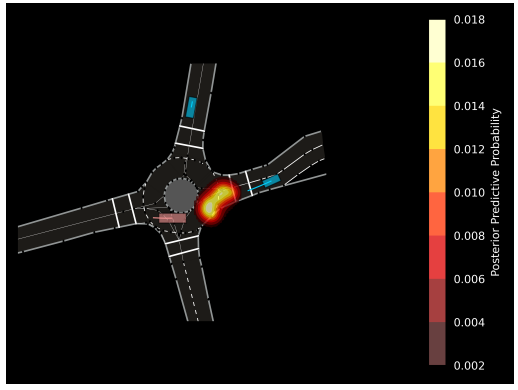


Figure: OOD case from `Roundabout_RW`

Table of Contents

1 Introduction

2 Methodology

3 Results

4 Conclusion

Conclusion

- **Summary**
  - 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.
- **Future Directions**
  - Model the full distributions for intermediate steps.
  - Propose an infinite mixture model for higher flexibility.
  - Enable multi-agent predictions from forward passing once.