Appendix

Table of Contents

A Theoretical Proofs

Definition 4 (Structural Hamming Distance (SHD))**.** *For any two DAGs $\mathcal{G}_1^C, \mathcal{G}_2^C$ with identical vertices set V, we define the following function SHD: $\mathcal{G} \times \mathcal{H} \rightarrow \mathbb{R}$,*

$$\begin{aligned} \textit{SHD}(\mathcal{G}_1^C, \mathcal{G}_2^C) &= \#\{(i,j) \in V^2 \mid \mathcal{G}_1^C \textit{ and } \mathcal{G}_2^C \textit{ have different edges } e_{ij}\} \\ &\triangleq \sum_{j \in V} |\boldsymbol{PA}_j(\mathcal{G}_1^C) - \boldsymbol{PA}_j(\mathcal{G}_2^C)| \end{aligned} \tag{6}$$

where $|\boldsymbol{PA}_j(\mathcal{G}_1^C) - \boldsymbol{PA}_j(\mathcal{G}_2^C)|$ is the number of the absolute difference in parental nodes for node j between causal graph $\mathcal{G}_1^C$ and $\mathcal{G}_2^C$.

Definition 5 (Nodes in Behavior Graph)**.** *Let $X_j = \left[V_j, \{E_{ij}\}_{i \in \{\boldsymbol{PA}_j(\mathcal{G}^C) \cup j\}}\right]$, where V_i is the node type of the j-th node, and $E_{\cdot i}$ is the arrows that point in the j-th node. All these components form the node X_j in the behavior graph.*

Definition 6 (Respect the graph)**.** *For any given behavior graph $\mathcal{G}^B$ with a specific causal graph $\mathcal{G}^C$, the transition model respects the graph if the distribution $p_\phi(\mathcal{G}^B|\mathcal{G}^C)$ can be factorized as:*

$$p(\mathcal{G}^B|\mathcal{G}^C) = \prod_{j \in [m]} p(X_j|\boldsymbol{PA}_j(\mathcal{G}^C)) \tag{7}$$

where m is the number of factorized nodes, and $\boldsymbol{PA}_j(\cdot)$ is for X_j's parents based on the causal graph.

Proposition 1 (CausalAF respects the graph)**.**

$$\begin{aligned} p_\phi(\mathcal{G}^B|\mathcal{G}^C) &= \prod_{j \in [m]} \Big[\underbrace{p_\phi(V_j|\boldsymbol{PA}_j(\mathcal{G}^C))}_{\textit{COM}} \underbrace{p_\phi(E_{jj}|V_j, \boldsymbol{PA}_j(\mathcal{G}^C)) \prod_{i \in \boldsymbol{PA}_j(\mathcal{G}^C)} p_\phi(E_{ij}|V_j, \boldsymbol{PA}_j(\mathcal{G}^C))}_{\textit{CVM}} \Big] \\ &= \prod_{j \in [m]} \Big[p_\phi(V_j, E_{jj}|\boldsymbol{PA}_j(\mathcal{G}^C)) \prod_{i \in \boldsymbol{PA}_j(\mathcal{G}^C)} p_\phi(E_{ij}|V_j, \boldsymbol{PA}_j(\mathcal{G}^C)) \Big] \\ &= \prod_{j \in [m]} p_\phi(V_j, \{E_{ij}\}_{i \in \{\boldsymbol{PA}_j(\mathcal{G}^C) \cup j\}}|\boldsymbol{PA}_j(\mathcal{G}^C)) \\ &= \prod_{j \in [m]} p_\phi(X_j|\boldsymbol{PA}_j(\mathcal{G}^C)) \end{aligned} \tag{8}$$

The node generation process of CausalAF combines two phases: firstly, we use COM to determine the generation order of the node, which prevents the generation of child nodes before their parent nodes. This COM can also be interpreted as a node ordering with topological sorting, therefore CausalAF should always respect the term $p(V_j|\boldsymbol{PA}_j(\mathcal{G}^C)), \forall j$ in Equation (8).

On the other hand, CVM is used to guarantee that the output of the autoregressive flow model uses proper structural information (i.e. the parents of the current node) to generate the self-loop edge as well as edges between new nodes and their parents accordingly, the CVM trick thus guarantees that CausalAF respects the term $p(E_{jj}|V_j, \boldsymbol{PA}_j(\mathcal{G}^C)) \prod_{i \in \boldsymbol{PA}_j(\mathcal{G}^C)} p(E_{ij}|V_j, \boldsymbol{PA}_j(\mathcal{G}^C)), \forall j$ in Equation (8).

Assumption 1 (Local Optimality)**.** *Let $\mathcal{G}^{C^*}$ be the ground truth causal graph, for any nodes X_j with its parental set $\boldsymbol{PA}_j(\mathcal{G}_1^C) \neq \boldsymbol{PA}_j(\mathcal{G}^{C^*})$. At convergence, CausalAF will have $\max_\phi p_\phi(V_j|\boldsymbol{PA}_j(\mathcal{G}^{C^*})) > \max_\phi p_\phi(V_j|\boldsymbol{PA}_j(\mathcal{G}_1^C))$.*

Assumption 2 (Local Monotonicity of Behavior Graph)**.** *For a single node X_j, its local monotonicity of likelihood means for any conditional set $\boldsymbol{PA}_j(\mathcal{G}_1^C), \boldsymbol{PA}_j(\mathcal{G}_2^C) \neq \boldsymbol{PA}_j(\mathcal{G}^C)$, if $|\boldsymbol{PA}_j(\mathcal{G}_1^C) - \boldsymbol{PA}_j(\mathcal{G}^C)| < |\boldsymbol{PA}_j(\mathcal{G}_2^C) - \boldsymbol{PA}_j(\mathcal{G}^C)|$, and $\exists\, v$, s.t. $\boldsymbol{PA}_j(\mathcal{G}_2^C) \cup v = \boldsymbol{PA}_j(\mathcal{G}_1^C)$, then $\max_\phi p_\phi(X_j|\boldsymbol{PA}_j(\mathcal{G}_1^C)) > \max_\phi p_\phi(X_j|\boldsymbol{PA}_j(\mathcal{G}_2^C))$*

Proof of Theorem 1. Given that $\mathcal{G}^B \sim p_\phi(\mathcal{G}^B|\mathcal{G}^C), \tau = \mathcal{E}(\mathcal{G}^B)$, by using the change of variable theorem, we have $\tau \sim p_\phi(\mathcal{E}^{-1}(\tau)|\mathcal{G}^C)|\det \frac{\partial \mathcal{E}^{-1}(\tau)}{\partial \tau}| \triangleq \hat{p}_\phi(\tau|\mathcal{G}^C)$.

The optimization process of CausalAF can be rewritten as below:

$$
\begin{aligned}
&\max_\phi \mathbb{E}_{\mathcal{G}^B \sim p_\phi(\mathcal{G}^B|\mathcal{G}^C)}[\mathbb{1}(D(\mathcal{E}(\mathcal{G}^B)]) < \epsilon) \\
&\quad = \max_\phi \mathbb{E}_{\hat{p}_\phi(\tau|\mathcal{G}^C)}[\mathbb{1}(D(\tau) < \epsilon)] \\
&\quad = \max_\phi \hat{p}_\phi(D(\tau) < \epsilon|\mathcal{G}^C) \\
&\quad = \max_\phi \hat{p}_\phi(\mathcal{G}^B \in \mathcal{A}|\mathcal{G}^C), \text{ where } \mathcal{A} = \{\mathcal{G}^B|D(\mathcal{E}(\mathcal{G}^B)) < \epsilon\}
\end{aligned}
\tag{9}
$$

Since the CausalAF respects the graph, as is shown in Proposition 1, for true CG $\mathcal{G}^{C^*}$ and another CG $\mathcal{G}_1^C \neq \mathcal{G}^{C^*}$. By applying the local monotonicity in the previous assumptions, when CausalAF converges, we will have

$$
\begin{aligned}
\hat{p}_\phi(\mathcal{G}^B \in \mathcal{A}|\mathcal{G}_1^C) &= \prod_j \hat{p}_\phi(X_j \in \mathcal{A}_j|\mathbf{PA}_j(\mathcal{G}_1^C)) \\
&= \prod_{\substack{\forall j, s.t. \\ \mathbf{PA}_j(\mathcal{G}_1^C) = \mathbf{PA}_j(\mathcal{G}^{C^*})}} \hat{p}_\phi(X_j \in \mathcal{A}_j|\mathbf{PA}_j(\mathcal{G}_1^C)) \prod_{\substack{\forall j, s.t. \\ \mathbf{PA}_j(\mathcal{G}_1^C) \neq \mathbf{PA}_j(\mathcal{G}^{C^*})}} \hat{p}_\phi(X_j \in \mathcal{A}_j|\mathbf{PA}_j(\mathcal{G}_1^C)) \\
&< \prod_{\substack{\forall j, s.t. \\ \mathbf{PA}_j(\mathcal{G}_1^C) = \mathbf{PA}_j(\mathcal{G}^{C^*})}} \hat{p}_\phi(X_j \in \mathcal{A}_j|\mathbf{PA}_j(\mathcal{G}^{C^*})) \prod_{\substack{\forall j, s.t. \\ \mathbf{PA}_j(\mathcal{G}_1^C) = \mathbf{PA}_j(\mathcal{G}^{C^*})}} \hat{p}_\phi(X_j \in \mathcal{A}_j|\mathbf{PA}_j(\mathcal{G}^{C^*})) \\
&= \prod_j \hat{p}_\phi(X_j \in \mathcal{A}_j|\mathbf{PA}_j(\mathcal{G}^{C^*})) \\
&= \hat{p}_\phi(\mathcal{G}^B \in \mathcal{A}|\mathcal{G}^{C^*})
\end{aligned}
\tag{10}
$$

Then we assume we have another Causal Graph $\mathcal{G}_2^C \neq \mathcal{G}_1^C$, if $SHD(\mathcal{G}_1^C, \mathcal{G}^{C^*}) < SHD(\mathcal{G}_2^C, \mathcal{G}^{C^*})$, and $\exists\, e$, s.t. $E_1^C \cup \{e\} = E_2^C$,

$$
\begin{aligned}
\hat{p}_\phi(\mathcal{G}^B \in \mathcal{A}|\mathcal{G}_2^C) &= \prod_j \hat{p}_\phi(X_j \in \mathcal{A}_j|\mathbf{PA}_j(\mathcal{G}_2^C)) \\
&= \prod_{\substack{\forall j, s.t. \\ \mathbf{PA}_j(\mathcal{G}_1^C)=\mathbf{PA}_j(\mathcal{G}_2^C)}} \hat{p}_\phi(X_j \in \mathcal{A}_j|\mathbf{PA}_j(\mathcal{G}_2^C)) \prod_{\substack{\forall j, s.t. \\ \mathbf{PA}_j(\mathcal{G}_1^C)\neq\mathbf{PA}_j(\mathcal{G}_2^C)}} \hat{p}_\phi(X_j \in \mathcal{A}_j|\mathbf{PA}_j(\mathcal{G}_2^C)) \\
&< \prod_{\substack{\forall j, s.t. \\ \mathbf{PA}_j(\mathcal{G}_1^C)=\mathbf{PA}_j(\mathcal{G}_2^C)}} \hat{p}_\phi(X_j \in \mathcal{A}_j|\mathbf{PA}_j(\mathcal{G}_1^C)) \prod_{\substack{\forall j, s.t. \\ \mathbf{PA}_j(\mathcal{G}_1^C)=\mathbf{PA}_j(\mathcal{G}_1^C)}} \hat{p}_\phi(X_j \in \mathcal{A}_j|\mathbf{PA}_j(\mathcal{G}_1^C)) \\
&= \prod_j \hat{p}_\phi(X_j \in \mathcal{A}_j|\mathbf{PA}_j(\mathcal{G}_1^C)) \\
&= \hat{p}_\phi(\mathcal{G}^B \in \mathcal{A}|\mathcal{G}_1^C)
\end{aligned} \tag{11}
$$

Based on the derivation above, we conclude that $\hat{p}_\phi(\mathcal{G}^B \in \mathcal{A}|\mathcal{G}_2^C) < \hat{p}_\phi(\mathcal{G}^B \in \mathcal{A}|\mathcal{G}_1^C) < \hat{p}_\phi(\mathcal{G}^B \in \mathcal{A}|\mathcal{G}^{C^*})$, which indicates that at convergence, the likelihood of collision samples converge with monotonicity guarantees:

$$p_\phi(D(\tau) < \epsilon \mid \mathcal{G}_2^C) < p_\phi(D(\tau) < \epsilon \mid \mathcal{G}_1^C) < p_\phi(D(\tau) < \epsilon \mid \mathcal{G}^{C^*}) \tag{12}$$

□

Table 3: Parameters of Environments

Parameter	Description	Value
S_{ego}	number of LiDAR sensor for ego vehicle	10
S_{other}	number of LiDAR sensor for other vehicle	0
S_{ped}	number of LiDAR sensor for pedestrian	6
M_{ego}	maximal range (m) of LiDAR for ego vehicle	200
M_{other}	maximal range (m) of LiDAR for other vehicle	200
M_{ped}	maximal range (m) of LiDAR for pedestrian	100
D_{ego}	braking factor of ego vehicle	0.1
D_{other}	braking factor of other vehicle	0.05
D_{ped}	braking factor of pedestrian	0.01
W_{ego}	shape size (width, length) of ego vehicle	[20, 40]
W_{other}	shape size (width, length) of ego vehicle	[20, 40]
W_{ped}	shape size (width, length) of ego vehicle	[15, 15]
V_{ego}	initial velocity of ego vehicle	18
V_{other}	initial velocity of other vehicle	18
V_{ped}	initial velocity of pedestrian	4
T_{max}	max number of step in one episode	100
C	collision threshold	20
Δ_t	step size of running	0.3

B Environment Details

B.1 Simulator

We conduct all of our experiments in a 2D traffic simulator, where vehicles and pedestrians are controlled by the Bicycle vehicle dynamics. The action is a two-dimensional continuous vector, containing the acceleration and steering. The ego vehicle is controlled by a constant velocity model and it will decelerate if its Radar detects some obstacles in front of it. All other objects are controlled by the scenario generation algorithm. The parameters of simulators and 3 environments are summarized in Table 3.

B.2 Definitions of Nodes and Edges in Causal Graph and Behavior Graph

In our experiments, we pre-define the types of nodes and types for Causal Graph and Behavior Graph, which is summarized in Table 4. Both of them share the same definition of node types. Causal Graph does not have the type of edges since it only describes the structure.

Table 4: Definitions of Nodes and Edges

Notation	Category	Description
n_N	Node type	empty node used as a placeholder in the vector
n_E	Node type	represents ego vehicle
n_V	Node type	represents non-ego vehicles
n_B	Node type	represents static objects in the scenario
n_P	Node type	represents pedestrian
e_N	Edge type	empty edge used as a placeholder in the vector
e_T	Edge type	the source node go toward the target node
e_S	Edge type	self-loop edge that does not rely on target node
e_p	Edge attribute	the initial 2D position of source node relative to target node
e_v	Edge attribute	the initial velocity of source node relative to target node
e_a	Edge attribute	the acceleration of source node relative to target node
e_s	Edge attribute	the shape size of the object in source node

C Model Training Details

Our model is implemented with PyTorch, using Adam as the optimizer. All experiments are conducted on NVIDIA GTX 1080Ti and Intel i9-9900K CPU@3.60GHz. We summarize the parameters of our model in Table 5. Note that the two variant models (Baseline and Baseline+COM) share the same parameters.

Table 5: Parameters of Environments

Parameter	Description	Value
E	episode number of REINFORCE	500
B	Batch size of REINFORCE	128
α	learning rate of REINFORCE	0.0001
T	sample temperature	0.5
m	maximal number of node	10
n	number of node type	5
n	number of node type	5
h_1	number of edge type	2
h_2	number of edge attribute	3
K	number of flow layer	2
d_h	dimension of hidden layer	128

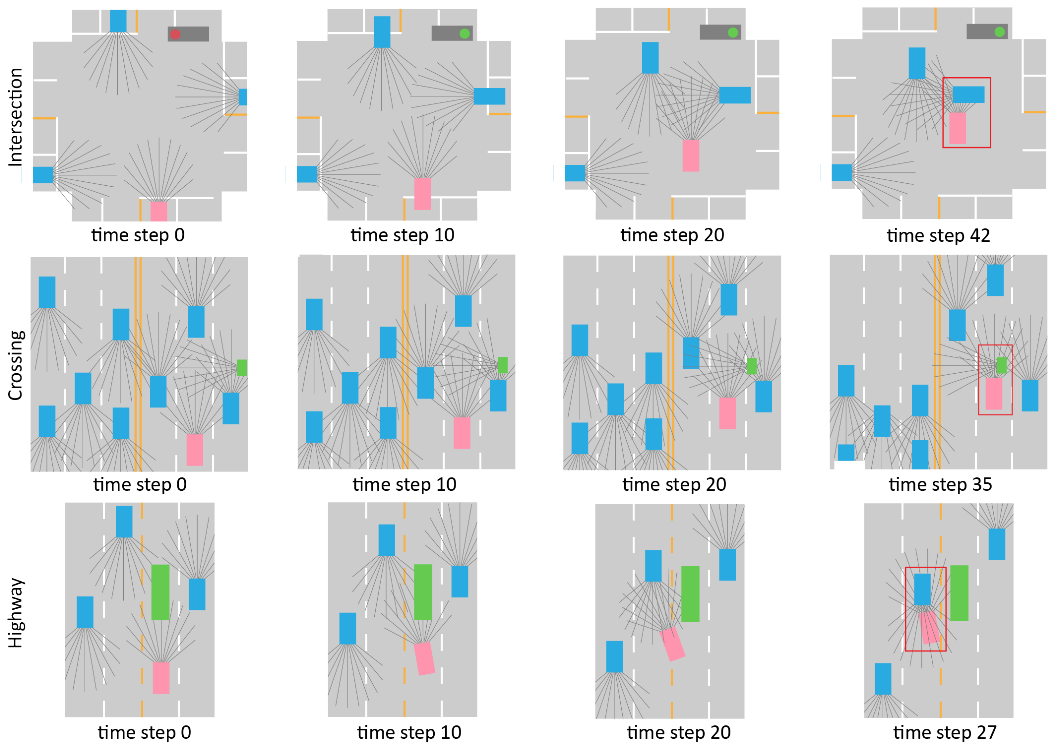

Figure 6: Screenshots of three generated scenarios in our simulator. The pink color represents the ego vehicle, the green color represents the pedestrian, and the blue color represents other vehicles. The red rectangle indicates the occurrence of a collision.

D More Experiment Results

D.1 Qualitative Results of Generated Scenarios

We show three qualitative results of generated safety-critical scenarios in Figure 6.

D.2 Diversity of Generated Scenarios

By injecting the causality into the generation process, we also restrict the space of generated scenario. Therefore, there usually exists a trade-off between the diversity and efficiency of generation. To analyze the diversity we lose by using the causal graph, we plot the variances of velocity and position of vehicles and pedestrians in Figure 7. We can see that the difference between the two models is very small, which indicates that the diversity of our CausalAF method is not decreased due to the injection of the causal graph.

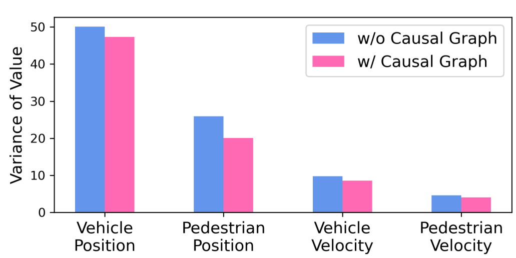

Figure 7: Variance of position and velocity of generated scenarios from two different models. One is with the causal graph and the other is without the causal graph.