Exploring Social Posterior Collapse in Variational Autoencoder for Interaction Modeling

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multi-agent behavior modeling and trajectory forecasting is crucial for safe navigation of autonomous agents in interactive scenarios. Variational Autoencoder (VAE) has been widely applied in multi-agent interaction modeling for its ability in generating diverse behavior and learning a low-dimensional representation for interacting systems. However, existing literature did not formally discuss if a VAE-based model can properly encode interaction into its latent space. In this work, we argue that one of typical formulations of VAEs in multi-agent modeling suffers from an issue we refer to as social posterior collapse, i.e., the model is prone to ignore social context when predicting the future trajectory of an agent. It could cause large prediction errors and poor generalization performance. We analyze the reason behind this under-explored phenomenon and propose several measures to tackle it. Afterwards, we implement the proposed framework and experiment on real-world datasets for multi-agent trajectory prediction. In particular, we propose a novel sparse graph attention message-passing (sparse-GAMP) layer which helps us detect social posterior collapse in our experiments. In the experiments, we verify that social posterior collapse indeed occurs. Also, the proposed measures are effective in alleviating the issue. As a result, the model attains better generalization performance when social context is informative for prediction.

1 Introduction

Modeling the behavior of intelligent agents is an important subject for autonomous systems. Safe operations of the autonomous agents require accurate prediction of other agents' future motion. In particular, social interaction among agents should be modeled, so that downstream planning and decision making modules can safely navigate the robots through interactive scenarios. Generative latent variable model is a popular modeling option for its ability in generating diverse and naturalistic behavior [1, 2, 3, 4]. In this work, we focus on one category of generative models, Variational Autoencoder (VAE) [5], which has been widely used in multi-agent behavior modeling and trajectory prediction [1, 4, 6, 7, 8, 9]. It is particularly interesting as it learns a low-dimensional representation of the original high-dimensional data.

However, VAEs do not necessarily learn a good representation of the data [10]. For instance, prior works in sequence modeling have found out that the model tends to ignore the latent variables if the decoder is overly powerful [11, 12, 13] (e.g. an autoregressive decoder). It results in a question on whether a VAE-based model can always learn a good representation for a multi-agent interacting system. It is a rather general question as researchers may look for different properties of the latent space, for instance, interpretability [14, 15] and multi-modality [4, 6]. In this work, we focus on a fundamental aspect of the question: Is interaction always properly modeled by the latent space? Formally, given a latent variable model of an interacting system, where each agent's behavior is

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

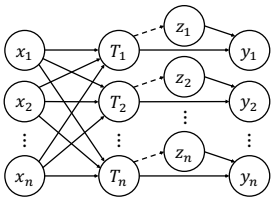

Figure 1: A latent variable model for interaction modeling. The dash edges from T_i to z_i apply only in the Conditional Variational Autoencoder (CVAE) setting.

governed by a latent variable, we are curious about if the VAE always learns to encode social context information into the latent variables.

It raises such a concern since VAE handles two distinct tasks in training and testing. At the training stage, it learns to reconstruct a datum instead of generation. For multi-agent interaction modeling, the sample for reconstruction is a set of trajectories of all the agents. Ideally, the model should learn to model the interaction among agents and jointly reconstruct the trajectories. However, there is no mechanism to prevent the model from reconstructing the trajectories in a separate manner. In fact, it could even be more efficient at the early stage of training, when the model has not learned an informative embedding of social context. We then end up with a model that models each agent's behavior separately without social context. which might suffer from over-estimated variance and large prediction error. More importantly, since the joint behavior of the agents is a consequence of their interactions, ignoring the causes may lead to poor generalization ability [15, 16]. We refer to this phenomenon as social posterior collapse. The issue has never been discussed in the literature, and it is necessary to initiate the study on such a crucial and fundamental issue.

In this work, we focus on a typical formulation of VAE for multi-agent interaction, which is abstracted from a wide range of existing literature [1, 4, 6, 8]. We first analyze social posterior collapse under this formulation and propose several measures to alleviate the issue. Afterwards, we study the issue under real-world settings with a realization of the abstract formulation we design. In particular, we propose a novel sparse graph attention message-passing (sparse-GAMP) layer and incorporate it into the model, which helps us detect and analyze social posterior collapse. In our experiments, we show that social posterior collapse occurs in real-world prediction tasks and the proposed measures are effective in alleviating the issue. We also evaluate how social posterior collapse affects the model performance on these tasks. The results suggest that the model without social posterior collapse can attain better generalization performance when social context is informative for prediction.

2 Social Posterior Collapse in Variational Autoencoder Interaction Models

2.1 A Latent Variable Model for Interaction Modeling

Given an interacting system with n agents, an interaction-aware trajectory prediction model takes the observed historical trajectories of all the agents, denoted by $\{\mathbf{x}_i\}_{i=1}^n$, as input and predicts the future trajectories of all the agents or a subset of enquired agents. We denote the collection of future trajectories by $\{\mathbf{y}_i\}_{i=1}^n$. In this work, we focus on the latent variable model illustrated in Fig. 1, which is abstracted from existing literature on VAEs for multi-agent behavior modeling [1, 4, 6, 8]. We model interaction by introducing a set of variables $\{\mathbf{T}_i\}_{i=1}^n$, which aggregates each agent's state and its observation of other agents. Formally, each $\mathbf{T}_i$ is modeled as a deterministic function of $\{\mathbf{x}_i\}_{i=1}^n$, i.e., $\mathbf{T}_i = f_i\left(\{\mathbf{x}_i\}_{i=1}^n\right)$. Afterwards, each agent makes decision based on the aggregated information over the predicted horizon. Latent variables $\{\mathbf{z}_i\}_{i=1}^n$ are introduced to model the inherent uncertainty in each agent's behavior. It should be noticed that interaction over the predicted horizon is not modeled explicitly in this formulation. Although it can be achieved by exchanging information between agents recurrently (e.g., social pooling in [17]), it is a common practice to avoid explicit modeling of future interaction in consideration of computational cost and memory [3, 8, 18].

In this work, we train the model as a VAE, where an encoder $q\left(\mathbf{z}|\mathbf{x},\mathbf{y}\right)$ is introduced to approximate the posterior distribution $p\left(\mathbf{z}|\mathbf{x},\mathbf{y}\right)$ for efficient sampling at the training stage[1]. The per-

[1]The vectors $\mathbf{x}, \mathbf{y}$ and $\mathbf{z}$ collect the corresponding variables for all n agents.

formance of variational inference is optimized if the KL-divergence between the posteriors, i.e. $D_{KL}\left[q\left(\mathbf{z}|\mathbf{x},\mathbf{y}\right)\|p\left(\mathbf{z}|\mathbf{x},\mathbf{y}\right)\right]$, is small [10]. To derive a better approximation, we can incorporate inductive bias based on the characteristics of the true posterior into the encoder function. We introduce the following simple proposition that guides our model design. Its proof can be found in Appx. A.

Proposition 1. *For any $i = 1, 2, ..., n$ and $j = 1, 2, ..., n$, 1) If $j \neq i$, $\mathbf{z}_i$ and $\mathbf{y}_j$ are conditionally independent given $\mathbf{x}$ and $\mathbf{y}_i$; 2) If $j \neq i$, $\mathbf{z}_i$ and $\mathbf{z}_j$ are conditionally independent given $\mathbf{x}$; 3) $\mathbf{z}_i$ and $\mathbf{x}_j$ are conditionally independent given $\mathbf{T}_i$.*

Following the proposition, we decompose the posterior distribution as $\prod_{i=1}^{n} p(\mathbf{z}_i|\mathbf{T}_i,\mathbf{y}_i)$. The decomposition suggests two insights. First, the encoder does not need to aggregate future context information when inferring the posterior distribution of $\mathbf{z}_i$ for each agent. Second, the historical contexts variable $\mathbf{T}_i$ is all we need for the historical information of the agent i. The encoder and decoder can share the same function to encode historical information.

2.2 Social Posterior Collapse

We then design a VAE model reflecting the characteristics of the true posterior discovered in Sec. 2.1. To simplify the problem, we further make the assumption of homogeneous agents. The model has three basic building blocks: 1) A function modeling the historical context, i.e., $\mathbf{T}_i = f_\theta\left(\mathbf{x}_i,\mathbf{x}\right)$; 2) A function decoding the distribution of the future trajectory $\mathbf{y}_i$ given $\mathbf{T}_i$ and $\mathbf{z}_i$, i.e., $p_\phi\left(\mathbf{y}_i|\mathbf{T}_i,\mathbf{z}_i\right)$; 3) A function approximating the posterior of $\mathbf{z}_i$ conditioned on $\mathbf{T}_i$ and $\mathbf{y}_i$, i.e., $q_\psi\left(\mathbf{z}_i|\mathbf{T}_i,\mathbf{y}_i\right)$. They build up the encoder and decoder of the VAE model as follows:

$$q_{\theta,\psi}\left(\mathbf{z}|\mathbf{x},\mathbf{y}\right) = \prod_{i=1}^{n} q_\psi\left(\mathbf{z}_i|f_\theta(\mathbf{x}_i,\mathbf{x}),\mathbf{y}_i\right), \qquad p_{\theta,\phi}\left(\mathbf{y}|\mathbf{x},\mathbf{z}\right) = \prod_{i=1}^{n} p_\phi\left(\mathbf{y}_i|f_\theta(\mathbf{x}_i,\mathbf{x}),\mathbf{z}_i\right). \quad (1)$$

We consider a continuous latent space and model q_ψ and p_ϕ as diagonal Gaussian distributions. The model is trained by maximizing the evidence lower-bound (ELBO):

$$\max_{\theta,\psi,\phi} \mathbb{E}_{\mathbf{x},\mathbf{y}\sim D}\left[\mathbb{E}_{\mathbf{z}\sim q_{\theta,\psi}(\mathbf{z}|\mathbf{x},\mathbf{y})}\left[\log p_{\theta,\phi}\left(\mathbf{y}|\mathbf{x},\mathbf{z}\right)\right] - \beta D_{KL}\left[q_{\theta,\psi}\left(\mathbf{z}|\mathbf{x},\mathbf{y}\right)\|p(\mathbf{z})\right]\right]. \quad (2)$$

However, we find a critical issue of this naive formulation during experiments, which is the social posterior collapse phenomenon mentioned before. With the help of the sparse graph attention mechanism introduced in Sec. 3, we find that the model is prone to ignore the other agents when reconstructing the future trajectory of one agent. Equivalently, the generative model collapses to the one with all the variables of other agents marginalized out.

We think the reason behind is similar to but different from the well-known phenomenon in VAE training—posterior collapse [10]. Due to the KL regularization term in ELBO, the variational posterior distribution is prone to collapse towards the prior. It is particularly likely to occur at the early stage of training, when the latent space is still uninformative [19]. In our case, the posterior of $\mathbf{z}_i$ collapses into $q_{\theta,\psi}(\mathbf{z}_i|\mathbf{x}_i,\mathbf{y}_i)$. It does minimize the KL-divergence: if both $q_{\theta,\psi}\left(\mathbf{z}_i|\mathbf{x}_i,\mathbf{y}_i\right)$ and $q_{\theta,\psi}(\mathbf{z}_i|\mathbf{x},\mathbf{y}_i)$ exactly approximate the true posteriors, then conditioning on more context information increases the expected value of KL regularization:

$$\mathbb{E}_D\left[D_{KL}\left[p\left(\mathbf{z}_i|\mathbf{x},\mathbf{y}_i\right)\|p(\mathbf{z}_i)\right]\right] = I(\mathbf{z}_i;\mathbf{x},\mathbf{y}_i) \geqslant I(\mathbf{z}_i;\mathbf{x}_i,\mathbf{y}_i) = \mathbb{E}_D\left[D_{KL}\left[p\left(\mathbf{z}_i|\mathbf{x}_i,\mathbf{y}_i\right)\|p(\mathbf{z}_i)\right]\right].$$

However, we argue that there is an additional factor contributing to the social posterior collapse problem, which makes it a unique phenomenon for interaction modeling. At the training stage, the goal is to reconstruct the future trajectories $\mathbf{y}$. The trajectory itself contains all the information needed for reconstruction. The history of the agent i provides complementary information such as current state and static environmental information. There is actually no explicit regulation in the current framework to prevent the model from extracting information solely from $\mathbf{x}_i$ and $\mathbf{y}_i$. In fact, it is an more efficient coding scheme, when the model has not learned an informative representation of interaction. Techniques such as KL annealing [13, 19] rely on various scheduling schemes of β to prevent KL vanishing, which have been shown effective in mitigating the posterior collapse problem. However, reducing β could merely privilege the model to gather more information from $\mathbf{y}_i$ which is consistent with the reconstruction objective. Therefore, we need to explore alternative solutions to tackle the social posterior collapse problem deliberately.

We start with changing the model into a conditional generative one [20, 21]. Additional edges $\{\mathbf{T}_i \rightarrow \mathbf{z}_i\}_{i=1}^{n}$ are added into the original graph, which are annotated with dash lines in Fig. 1.

We follow the practice in [20] and formulate the model as a Conditional Variational Autoencoder (CVAE). It is straightforward to verify that the conclusion of Proposition 2.1 still applies. We still model the encoder and decoder as in Eqn. (1). The CVAE framework introduces an additional module —a function approximating the conditional prior $p_\eta\left(\mathbf{z}_i|\mathbf{T}_i\right)$, which becomes $p_{\theta,\eta}\left(\mathbf{z}_i|\mathbf{x}\right)$ after incorporating $f_\theta(\mathbf{x})$. The objective then becomes:

$$\mathcal{L}(\theta,\psi,\phi,\eta)=\mathbb{E}_{\mathbf{x},\mathbf{y}\sim D}\left[\mathbb{E}_{\mathbf{z}\sim q_{\theta,\psi}(\mathbf{z}|\mathbf{x},\mathbf{y})}\left[\log p_{\theta,\phi}\left(\mathbf{y}|\mathbf{x},\mathbf{z}\right)\right]-\beta D_{KL}\left[q_{\theta,\psi}\left(\mathbf{z}|\mathbf{x},\mathbf{y}\right)\|p_{\theta,\eta}(\mathbf{z}|\mathbf{x})\right]\right].$$

Compared to Eqn. (2), we no longer penalize the encoder for aggregating context information but only restrict the information encoded from future trajectories. Therefore, the encoder does not need to ignore context information to fulfill the information bottleneck. However, the model still lacks a mechanism to deliberately encourage context information encoding. To achieve the goal, we propose to incorporate an auxiliary task into the training scheme. Concretely, we introduce another module $\mathbf{y}_i=g_\zeta\left(\mathbf{T}_i,\mathbf{z}_i\right)$. Composing g_ζ with f_θ gives us another decoder $\mathbf{y}=h_{\theta,\zeta}(\mathbf{x},\mathbf{z})$. The difference is that the latent variables are always sampled from $p_\eta\left(\mathbf{z}_i|\mathbf{T}_i\right)$. The auxiliary task is training this trajectory decoder which shares the same context encoder and conditional prior with the CVAE. Because the auxiliary model does not have access to the ground-truth future trajectory, it needs to utilize context information for accurate prediction. Consequently, it encourages the model to encode context information into $\mathbf{T}$ and $\mathbf{z}$. We use mean squared error (MSE) loss as the objective function. The overall objective minimizes a weighted sum of the two objective functions.

The training scheme looks similar to the one in [20], where they trained a Gaussian stochastic neural network (GSNN) together with the CVAE model. However, ours is different from theirs in some major aspects. The GSNN model shares the same decoder with the CVAE model. The main motivation to incorporate another learning task to directly optimize generation procedure during training. In constrast, our auxiliary prediction model has a separate trajectory decoder. It is because we do not want the auxiliary task to interfere the decoding procedure of the CVAE model. We find that sharing the decoder leads to less diversity in trajectory generation, which is not desirable.

3 Sparse Graph Attention Message-Passing Layer

Before jumping to introducing the specific model, we develop for real-world prediction tasks, we would like to present a novel sparse graph attention message-passing (sparse-GAMP) layer, which helps us detect and analyze the social posterior collapse phenomenon.

3.1 Sparse Graph Attention Mechanism

Our sparse-GAMP layer incorporates α-entmax [22] as the graph attention mechanism. We are particularly interested in the 1.5-entmax variant, which is smooth and can be exactly computed. It is also easy to implement on GPUs using existing libraries (e.g., PyTorch [23]). Concretely, the 1.5-entmax function maps a d-dimensional input $\mathbf{s}\in\mathbb{R}^d$ into $\mathbf{p}\in\Delta^d=\left\{\mathbb{R}^d:\mathbf{p}\geqslant 0,\|\mathbf{p}\|_1=1\right\}$ as $\mathbf{p}=\left[\mathbf{s}/2-\tau\mathbf{1}\right]_+^2$, where τ is a unique threshold value computed using $\mathbf{s}$. Upon the theoretical results in [22], we derive an insightful proposition which makes 1.5-entmax a merited option in our framework. The proof of the proposition can be found in Appx. B.

Proposition 2. *Let $s_{[d]}\leqslant\cdots\leqslant s_{[1]}$ denote the sorted coordinates of $\mathbf{s}$. Define the top-ρ mean, unnormalized variance, and induced threshold for $\rho\in\{1,...,d\}$ as*

$$M_{\mathbf{s}}(\rho)=\frac{1}{\rho}\sum_{j=1}^{\rho}s_{[j]},\ S_{\mathbf{s}}(\rho)=\sum_{j=1}^{\rho}\left(s_{[j]}-M_{\mathbf{s}}(\rho)\right)^2,\ \tau_{\mathbf{s}}(\rho)=\begin{cases}M_{\mathbf{s}}(\rho)-\sqrt{\frac{1-S_{\mathbf{s}}(\rho)}{\rho}}, & S_{\mathbf{s}}(\rho)\leqslant 1,\\ +\infty, & \textit{otherwise.}\end{cases}$$

Let $\mathbf{s}'\in\mathbb{R}^{d+1}$ satisfy $s'_i=s_i$ for $i\leqslant d$, and define $\mathbf{p}=1.5$-entmax$(\mathbf{s})$ and $\mathbf{p}'=1.5$-entmax$(\mathbf{s}')$.

Then we have: 1) If $\frac{s'_{d+1}}{2}\leqslant\frac{s_{[d]}}{2}-1$, then $p'_i=p_i$ for $i=1,...,d$ and $p'_{d+1}=0$; 2) If $p_i>0$ for $i=1,..,d$, then $p'_i=p_i$ for $i=1,2,...,d$ and $p'_{d+1}=0$ iff $s'_{d+1}\leqslant 2\tau_{\mathbf{s}/2}(d)$.

The first statement of the proposition provides a sufficient condition for augmenting an input vector without affecting its original attention values. It is a very useful when applying 1.5-entmax to graph attention. Unlike typical neural network models, graph neural networks (GNN) operate on graphs

whose size may vary over different samples. Meanwhile, nodes within the same graph may have different number of incoming edges. Therefore, the 1.5-entmax function has inputs of varying dimension even within the same batch of training samples, which makes it inefficient to compute using available primitives. The proposition suggests a simple solution to this problem. Given $\{\mathbf{s}_j\}_{j=1}^m$ with $\mathbf{s}_j \in \mathbb{R}^{d_j}$, we can compute a dummy value as $\min_{j\in\{1,\dots,m\},i\in\{1,\dots,d_j\}} 2s_{j,i} - 2$. We can augment the input vectors with this dummy value to transform them into a matrix in $\mathbb{R}^{m\times d^*}$, where d^* is the largest value of d_j. The dummy elements will not affect the attention computation and existing primitives based on matrix computation can be directly used.

The second statement implies that the activated coordinates determine a unique threshold value for augmented coordinates. This property is particularly useful when modeling interactions with large number of agents (e.g., dense traffic scenes). It ensures that the effects of interacting agents will not be diluted by the irrelevant agents, which could potentially improve the robustness of the model [24]. In this work, we mainly utilize the interpretability of the sparse graph attention, but we will investigate its application in generalization in future work.

3.2 Sparse-GAMP

To obtain the sparse-GAMP layer, we combine the sparse graph attention with a message-passing GNN [25, 26]. Given a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertices $v \in \mathcal{V}$ and edges $e = (v, v') \in \mathcal{E}$. we define the sparse-GAMP layer as a composition of a node-to-edge message-passing step $v \to e$ and a edge-to-node message-passing step $e \to v$:

$$\begin{aligned} v \to e:\ & \mathbf{h}_{(i,j)} = f_e\left(\left[\mathbf{h}_i, \mathbf{h}_j, \mathbf{u}_{(i,j)}\right]\right), \\ e \to v:\ & \hat{\mathbf{h}}_j = \sum_{i\in\mathcal{N}_j} w_{(i,j)}\mathbf{h}_{(i,j)}, \ \text{ where } \mathbf{w}_j = \mathcal{G}\text{-entmax}\left(\left\{\mathbf{h}_{(i,j)}\right\}_{i\in\mathcal{N}_j}\right), \end{aligned} \tag{3}$$

where $\mathbf{h}_i$ is the embedding of node v_i produced by upstream modules in the pipeline, and $\mathbf{u}_{(i,j)}$ summarizes auxiliary edge features of edge $e_{(i,j)}$ (e.g. edge type). $\mathcal{N}_j$ denotes the set of indices of neighbor nodes connected by an incoming edge and $[\cdot,\cdot]$ denotes concatenation of vectors. The function f_e is a neural network encodes edge embedding for message passing. $\mathcal{G}$-entmax is the sparse graph attention function empowered by 1.5-entmax. It takes the edge embeddings of all the incoming edges of a given node as input, and computes the weights assigned to those edges.

4 Social-CVAE

In this section, we design a realization of the abstract framework studied in Sec. 2 for real-world trajectory prediction tasks. We choose to design the model with GNNs, which enable a flexible graph representation of data. Several GNN-based approaches achieved the state-of-the-art performance on trajectory prediction task [27, 28, 29]. The resulting model is depicted in Fig. 2, which we refer to as social-CVAE. During training, we first encode the historical and future trajectories of all the agents using a Gated Recurrent Unit (GRU) [30] network. For vehicle trajectory prediction tasks, we incorporate the map information in a manner similar to [31]. Different from [31] where road boundaries are modeled as nodes in the scene graph, we adopt the representation in [32] where the map is denoted as a graph consisting of lanelet nodes, i.e., drivable road segments. For each lanelet, it is represented by its left and right boundaries composed by two sequences of points. We use another GRU network to encode them. Although not utilized in this work, the lanelet representation allows us to combine routing information in a natural manner as in [29]. We will investigate it in future work.

To encode the historical context information, we construct a graph using the embeddings of historical trajectories and lanelets if available. Each pair of agent nodes has a bidirectional edge connecting each other. For each lanelet node, we add edges connecting it to all the agent nodes. If the map information is not available, we add self-edges for all the agents to enable direct self-encoding channels. One sparse-GAMP layer is applied to encode historical context for all the agents. If social posterior collapse occurs, the agent-to-agent edges, except for self-edges, are more likely to receive zero attention weights thanks to the sparse attention. Therefore, we can use the percentage of unattended agent-to-agent edges as a metric to detect social posterior collapse. It is a more objective metric than looking into the quantitative magnitude of attention weights. We can assert that an agent node does not contribute to the output if its attention is strictly zero. However, we need to be more careful

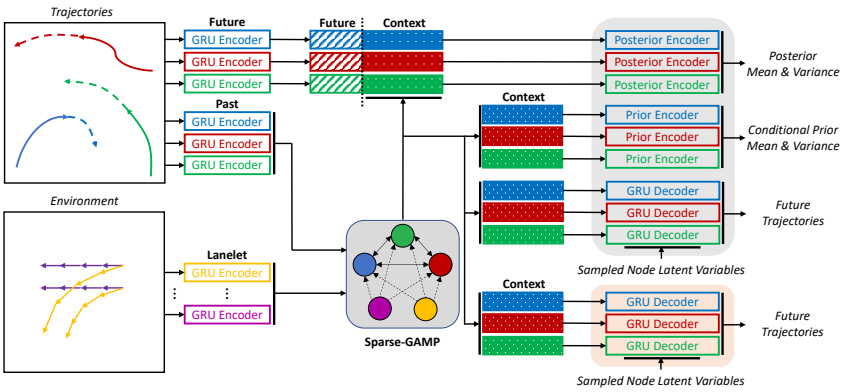

Figure 2: Social-CVAE architecture. We adopt a graph representation and use sparse-GAMP to encode context information. The modules within the orange box are the auxiliary decoders.

when comparing the importance of two agents based on non-zero attention weights [33, 34]. For the same reason, we do not consider techniques such as multiple message-passing layers [31, 28, 29] or separating self-edges from aggregation [8]. Even though they can potentially boost the performance, it might become inconclusive to analyze the model based on the attention map. After computing the historical context, we use a MLP to model the conditional Gaussian prior $p_\eta\left(\mathbf{z}_i|\mathbf{T}_i\right)$. To model the variational posterior $q_\psi\left(\mathbf{z}_i|\mathbf{T}_i,\mathbf{y}_i\right)$, we concatenate the embedding of future trajectory with the historical context for each agent, and use another MLP to output the posterior mean and variance. For the decoder, we use another GRU to decode the future trajectory for each agent separately. The auxiliary decoder shares the same structure but always samples the latent variables from the conditional prior regardless of training or testing.

5 Experiments

In this section, we report the experimental results on two trajectory prediction tasks. The main purpose is not achieving state-of-the-art performance, but to study the social posterior collapse problem. We compare social-CVAE with two variants of it: 1) A model without the auxiliary task; 2) A model without the auxiliary task or the conditional prior. They correspond to the vanilla CVAE and VAE formulations discussed in Sec.2. We are curious about whether the proposed measures can alleviate social posterior collapse. We will briefly introduce the experiment settings and present the main results. Please refer to Appx. C-F for more details, visualization, and additional experiments.

5.1 Vehicle Trajectory Prediction

The first task is the vehicle trajectory prediction problem, where we are asked to predict the future trajectory of one target vehicle given the historical trajectories of itself and other surrounding vehicles. We train the models on two datasets, INTERACTION Dataset [35] and Argoverse Motion Forecasting dataset [36]. To evaluate the prediction performance, we use the standard minimum Average Displacement Error (minADE) and minimum Final Displacement Error (minFDE) over K sampled trajectories as metrics. And we follow [36] to define minADE as ADE of the trajectory with minimum FDE. Additionally, we define a unique metric, Agent Ratio (AR), to study social posterior collapse:

$$\mathrm{AR} = \frac{\sum_{i=1,i\neq j}^{n} \mathbf{1}(\omega_{(i,j)} \neq 0)}{n-1},$$

where the agent j refers to the predicted target vehicle and $w_{(i,j)}$ is the attention weight assigned to the edge from the agent i to the target vehicle. It equals to the percentage of surrounding vehicles which receive non-zero attention. A value close to zero implies that the model ignores the majority of the surrounding vehicles, which is a sign of social posterior collapse.

INTERACTION Dataset. The INTERACTION dataset provides three categories of driving scenarios: intersection (IS), roundabout (RA) and highway merging (HM). For each experiment, we ran five

Table 1: INTERACTION Dataset Validation and Test Results

Scene	Model	AR(%)	K = 1		K = 6	
			minFDE	minADE	minFDE	minADE
IS	VAE	1.85 ± 4.02	1.32 ± 0.03	0.42 ± 0.01	0.73 ± 0.01	0.26 ± 0.01
	CVAE	0.18 ± 0.36	$\mathbf{1.27 \pm 0.02}$	$\mathbf{0.41 \pm 0.01}$	$\mathbf{0.68 \pm 0.02}$	$\mathbf{0.24 \pm 0.01}$
	Ours	$\mathbf{15.8 \pm 10.4}$	$\mathbf{1.27 \pm 0.02}$	$\mathbf{0.41 \pm 0.01}$	0.70 ± 0.02	0.25 ± 0.01
RA	VAE	0.05 ± 0.04	1.34 ± 0.02	0.42 ± 0.01	0.76 ± 0.01	0.26 ± 0.01
	CVAE	13.4 ± 14.9	1.32 ± 0.01	0.42 ± 0.01	0.73 ± 0.02	$\mathbf{0.25 \pm 0.01}$
	Ours	$\mathbf{19.5 \pm 15.4}$	$\mathbf{1.29 \pm 0.03}$	0.42 ± 0.01	$\mathbf{0.72 \pm 0.01}$	0.26 ± 0.01
HM	VAE	8.68 ± 8.36	0.87 ± 0.08	0.29 ± 0.03	0.42 ± 0.04	0.16 ± 0.01
	CVAE	5.42 ± 10.2	0.83 ± 0.03	0.28 ± 0.03	0.40 ± 0.05	0.16 ± 0.02
	Ours	$\mathbf{15.5 \pm 8.13}$	$\mathbf{0.65 \pm 0.09}$	$\mathbf{0.22 \pm 0.02}$	$\mathbf{0.32 \pm 0.04}$	$\mathbf{0.13 \pm 0.01}$

Track	Model	AR(%)	K = 6			K = 50		
			ADE	FDE	MoN	ADE	FDE	MoN
R	VAE	0.35 ± 5.37	0.5685	1.7573	0.2238	0.5712	1.7709	0.1036
	CVAE	0.07 ± 1.43	0.5323	1.6425	0.2144	0.5410	1.6725	**0.0983**
	Ours*	0.88 ± 2.49	0.5157	1.5823	0.2195	0.5158	1.5823	0.1106
	Ours	$\mathbf{24.8 \pm 20.2}$	**0.4665**	**1.4174**	**0.2011**	**0.4662**	**1.4187**	0.1050
G	VAE	0.02 ± 1.27	1.3428	3.8542	0.8193	1.3436	3.8564	0.5181
	CVAE	0.05 ± 2.20	1.4517	4.2179	0.8743	1.3824	3.9972	0.5676
	Ours*	0.37 ± 2.74	1.1615	3.3927	0.6811	1.1617	3.3939	**0.4218**
	Ours	$\mathbf{15.2 \pm 14.7}$	**0.9205**	**2.7049**	**0.6075**	**0.9186**	**2.6969**	0.4891

Ours* refers to an instance of social-CVAE that still suffers from the social posterior collapse problem.

trials to account for the randomness in initialization. We then evaluated them on the validation sets and computed the mean and standard deviation of the evaluated metrics over all the trials. The best models were selected for testing on the regular (R) and generalization (G) tracks of the INTERPRET Challenge [2]. The results are summarized in Table 1. The CVAE variants have AR values similar to the VAE variants on IS and HM. It is consistent with our argument that merely changing the model to a conditional one is insufficient. In contrast, our social-CVAE model consistently attains high AR values. However, the CVAE variant achieves similar prediction performance as ours on the validation sets of IS and RA, even if the CVAE variant suffers from the social posterior collapse problem. It is because the driving behavior within intersections and roundabouts depends highly on locations. When the map is less informative (e.g. validation set in HM) or a novel scenario is encountered (e.g. generalization track), our social-CVAE model which does not have social posterior collapse outperforms the other variants. Particularly, we compare it with another instance that attains a AR value close to zero even with the auxiliary task. Their test performances, especially on the generalization track, are quite different. It shows it is mainly the social context that improves the prediction performance. Even if the auxiliary task itself also contributes, the improvement is not significant, especially in novel scenes, unless it indeed prevents social posterior collapse.

Argoverse Dataset. Similar to the INTERACTION dataset, we trained five models with random initialization for each case and evaluated them on the validation sets. The best social-CVAE instance was selected for submission to the Argoverse Motion Forecasting Challenge. Although achieving state-of-the-art performance is not our objective, we still report the testing result in Appx. F and compare it with the other models on the leaderboard, in order to provide the audience a complete picture of the model. Here, we mainly focus on the validation results in Table 2. The conclusion is consistent. The difference is that the social-CVAE model outperforms the others even on the validation set. It is because the validation set of the Argoverse dataset is collected in different areas of the cities, which is more analogous to the generalization track of the INTERPRET Challenge.

[2]The challenge adopts a different group of evaluation metrics, please check their website for the formal definitions: `http://challenge.interaction-dataset.com/prediction-challenge/intro`

Table 2: Argoverse Dataset Validation Results

Model	AR(%)	K = 1		K = 6	
		minFDE	minADE	minFDE	minADE
VAE	1.27 ± 1.00	4.17 ± 0.17	1.86 ± 0.07	2.33 ± 0.06	1.30 ± 0.04
CVAE	0.40 ± 0.31	3.85 ± 0.01	1.73 ± 0.01	2.09 ± 0.02	1.19 ± 0.01
Ours	$\mathbf{28.5 \pm 15.7}$	$\mathbf{3.52 \pm 0.03}$	$\mathbf{1.59 \pm 0.02}$	$\mathbf{1.98 \pm 0.02}$	$\mathbf{1.15 \pm 0.01}$

5.2 Pedestrian Trajectory Prediction

The second task is the pedestrian trajectory prediction problem, where we are asked to forecast the future trajectories of all the pedestrians in each scenario. We trained the models on the well-established ETH/UCY dataset, which combines two datasets ETH [37] and UCY [38]. Following prior works [2, 39, 4, 7, 9], we adopt the leave-one-out evaluation protocol and do not use any environmental information. All the prediction metrics are computed with $K = 20$ samples. Similar to the vehicle case, we ran five trials for each experiment. The results are summarized in Table 3, where we report the testing results for each group and their average over all the groups. The comparison between our model and other methods can be found in Appx. F. In Table 3, we see that changing to CVAE formulation leads to a significant boost in terms of AR, implying that the model gives more attention to surrounding agents. However, neither changing to CVAE nor the auxiliary task improve prediction performance. It is because historical social context is less informative for pedestrian trajectory prediction. The ETH/UCY dataset is collected in unconstrained environments with few objects. Also, compared to vehicles, pedestrians have fewer physical constraints in 2D motion, resulting in shorter temporal dependency in their movement. It is consistent with the results in [9], where the authors proposed a transformer-based model that achieves the current state-of-the-art performance on ETH/UCY. The visualized attention maps in [9] show that social context in nearby timesteps is more important to their prediction model, whereas the historical trajectories of other agents always receive low attention weights. Therefore, even if the auxiliary prediction task encourages our model to encode historical social context, it does not lead to either larger AR value or better prediction performance. It suggests that the simplified latent variable model studied in this work might not be sufficient to model pedestrian motion. Explicit future interaction should be considered and an autoregressive decoder as the one in [9] should be used to account for it.

6 Discussion

Limitations. As the first study on social posterior collapse, we are not able to explore every aspect of this subject. There are many possible factors that could contribute to this phenomenon. Our experimental results can only speak for the particular model we develop and the tasks we study. The occurrence of social posterior collapse and its effect to the overall performance may vary across different problems and different model architectures. For instance, our findings from the pedestrian trajectory prediction task are quite different from its vehicle counterpart. The format of the graph representation, especially how the environmental information is incorporated, also matters. However, we do demonstrate it in an ablation study (Appx. E) that social posterior collapse is not unique to our sparse-GAMP module. Even if we switch to other aggregation functions, we can still find evidence suggesting its occurrence. Nevertheless, our sparse graph attention function, $\mathcal{G}$-entmax, is a convenient and flexible toolkit that allows us to monitor and analyze social posterior collapse without compromising performance. It will be helpful when we explore social posterior collapse in other interacting systems with alternative model architectures in the future.

Connections to Related Works. We would like to wrap up our discussion with a glance at those prior works on VAE-based multi-agent behavior modeling. We are curious about if any elements in their models have implicitly tackled this issue. However, we would like to emphasize that it is still necessary to explicitly study social posterior collapse under their settings in the future, which could provide a useful guidance to avoid social posterior collapse in model design. Due to the space limit, we only discuss works that used techniques we think are potentially related to social posterior collapse, especially those provided evidence that interacting behaviors were properly modeled (e.g., attention maps or visualized prediction results in interactive scenarios).

Table 3: ETH/UCY Dataset Leave-one-out Testing Results

Model	ETH			HOTEL		
	AR(%)	minFDE	minADE	AR(%)	minFDE	minADE
VAE	74.0 ± 7.33	0.98 ± 0.07	0.63 ± 0.03	29.0 ± 28.1	0.28 ± 0.01	0.19 ± 0.01
CVAE	$\mathbf{90.9 \pm 6.38}$	$\mathbf{0.94 \pm 0.10}$	$\mathbf{0.61 \pm 0.05}$	69.8 ± 41.7	$\mathbf{0.27 \pm 0.01}$	$\mathbf{0.18 \pm 0.01}$
Ours	83.2 ± 12.6	1.08 ± 0.10	0.68 ± 0.04	$\mathbf{72.9 \pm 21.5}$	$\mathbf{0.27 \pm 0.02}$	$\mathbf{0.18 \pm 0.01}$
Model	ZARA1			ZARA2		
	AR(%)	minFDE	minADE	AR(%)	minFDE	minADE
VAE	40.3 ± 22.8	$\mathbf{0.38 \pm 0.01}$	$\mathbf{0.22 \pm 0.01}$	30.1 ± 17.1	$\mathbf{0.32 \pm 0.01}$	$\mathbf{0.18 \pm 0.01}$
CVAE	$\mathbf{89.4 \pm 7.65}$	0.39 ± 0.01	$\mathbf{0.22 \pm 0.01}$	$\mathbf{64.2 \pm 25.6}$	0.35 ± 0.01	0.19 ± 0.01
Ours	61.9 ± 5.71	$\mathbf{0.38 \pm 0.01}$	$\mathbf{0.22 \pm 0.01}$	58.6 ± 15.7	0.37 ± 0.02	0.20 ± 0.01
Model	UNIV			Average		
	AR(%)	minFDE	minADE	AR(%)	minFDE	minADE
VAE	0.06 ± 0.09	$\mathbf{0.55 \pm 0.01}$	$\mathbf{0.31 \pm 0.01}$	34.7 ± 29.4	$\mathbf{0.50 \pm 0.27}$	$\mathbf{0.30 \pm 0.17}$
CVAE	$\mathbf{41.2 \pm 25.1}$	0.63 ± 0.05	0.35 ± 0.02	$\mathbf{71.1 \pm 29.5}$	0.51 ± 0.25	0.31 ± 0.17
Ours	36.3 ± 14.4	0.63 ± 0.02	0.35 ± 0.01	62.6 ± 21.0	0.54 ± 0.29	0.33 ± 0.19

Trajectron [6] and Trajectron++ [4] which are formulated as CVAEs establish the previous state-of-the-art results on ETH/UCY. Different from ours, they adopted a discrete latent space to account for the multi-modality in human behavior. They did not examine how social context is utilized by their models. However, we think that a discrete latent space could potentially alleviate the social posterior collapse issue. Compared to a continuous random variable, a discrete one can only encode a limited amount of information. It prevents the model from bypassing social context since a discrete latent variable might not be sufficient to encode all the information of a long trajectory. PECNet [7] is another CVAE-based trajectory prediction model. Instead of conditioning on the entire future trajectory, they proposed the Endpoint VAE where the posterior latent variables condition on the endpoints only. We think it could also be a potential solution as it also limits the amount of information the latent space can encode from the future trajectory. In [1], Suo et al. proposed a multi-agent behavior model for traffic simulation under the CVAE framework. They demonstrated that their method was able to simulate complex and realistic interactive behaviors. In particular, they augmented ELBO with a commonsense objective that regularizes the model from synthesizing undesired interactions (e.g. collisions). We think it plays a similar role as our auxiliary prediction task in terms of enforcing social context encoding. The last work we would like to discuss is the AgentFormer model [9] mentioned in Sec. 5.2. Their results suggest that the combination of autoregressive decoder and future social context encoder could be more effective in interaction modeling, especially for systems without long-term dependency (e.g. pedestrians). However, as we have mentioned before, the autoregressive decoder introduces another issue if it is overly powerful.

7 Conclusion

In this work, we point out a under-explored issue, which we refer to as social posterior collapse, in the application of VAEs in multi-agent modeling. We argue that one of the commonly adopted formulations of VAEs in multi-agent modeling is prone to ignore social context when predicting the future trajectory of an agent. We analyze the reason behind and propose several measures to alleviate social posterior collapse. Afterwards, we design a GNN-based realization of the general framework that incorporates the proposed measures, which we refer to as social-CVAE. In particular, social-CVAE incorporates a novel sparse-GAMP layer which helps us detect and analyze social posterior collapse. In our experiments, we show that social posterior collapse occurs in real-world trajectory prediction problems and the proposed measures are effective in alleviating the issue. Also, the experimental results imply that social posterior collapse could cause poor generalization performance in novel scenarios if the future movement of the agents indeed depends on their historical social context. In the future, we will utilize the toolkit developed in this work to explore social posterior collapse in a broader range of model architectures and interacting systems.

References

[1] S. Suo, S. Regalado, S. Casas, and R. Urtasun, "Trafficsim: Learning to simulate realistic multi-agent behaviors," *arXiv preprint arXiv:2101.06557*, 2021.

[2] 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.

[3] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. Reid, H. Rezatofighi, and S. Savarese, "Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks," in *Advances in Neural Information Processing Systems*, 2019.

[4] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," *arXiv preprint arXiv:2001.03093*, 2020.

[5] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[6] B. Ivanovic and M. Pavone, "The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2375–2384, 2019.

[7] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, and A. Gaidon, "It is not the journey but the destination: Endpoint conditioned trajectory prediction," in *European Conference on Computer Vision*, pp. 759–776, Springer, 2020.

[8] J. Li, H. Ma, Z. Zhang, and M. Tomizuka, "Social-wagdat: Interaction-aware trajectory prediction via wasserstein graph double-attention network," *arXiv preprint arXiv:2002.06241*, 2020.

[9] Y. Yuan, X. Weng, Y. Ou, and K. Kitani, "Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting," *arXiv preprint arXiv:2103.14023*, 2021.

[10] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel, "Variational lossy autoencoder," *arXiv preprint arXiv:1611.02731*, 2016.

[11] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," *arXiv preprint arXiv:1511.06349*, 2015.

[12] M. Fraccaro, S. K. Sønderby, U. Paquet, and O. Winther, "Sequential neural models with stochastic layers," *arXiv preprint arXiv:1605.07571*, 2016.

[13] I. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

[14] Y. Hu, W. Zhan, L. Sun, and M. Tomizuka, "Multi-modal probabilistic prediction of interactive behavior via an interpretable model," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, pp. 557–563, IEEE, 2019.

[15] C. Tang, N. Srishankar, S. Martin, and M. Tomizuka, "Grounded relational inference: Domain knowledge driven explainable autonomous driving," *arXiv preprint arXiv:2102.11905*, 2021.

[16] P. de Haan, D. Jayaraman, and S. Levine, "Causal confusion in imitation learning," *arXiv preprint arXiv:1905.11979*, 2019.

[17] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] S. Khandelwal, W. Qi, J. Singh, A. Hartnett, and D. Ramanan, "What-if motion prediction for autonomous driving," *arXiv preprint arXiv:2008.10587*, 2020.

[19] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin, "Cyclical annealing schedule: A simple approach to mitigating KL vanishing," in *NAACL*, 2019.

[20] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, pp. 3483–3491, 2015.

[21] B. Ivanovic, K. Leung, E. Schmerling, and M. Pavone, "Multimodal deep generative models for trajectory prediction: A conditional variational autoencoder approach," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 295–302, 2020.

[22] B. Peters, V. Niculae, and A. F. Martins, "Sparse sequence-to-sequence models," *arXiv preprint arXiv:1905.05702*, 2019.

[23] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.

[24] A. Goyal, A. Lamb, J. Hoffmann, S. Sodhani, S. Levine, Y. Bengio, and B. Schölkopf, "Recurrent independent mechanisms," *arXiv preprint arXiv:1909.10893*, 2019.

[25] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International Conference on Machine Learning*, pp. 1263–1272, PMLR, 2017.

[26] T. Kipf, E. Fetaya, K.-C. Wang, M. Welling, and R. Zemel, "Neural relational inference for interacting systems," in *International Conference on Machine Learning*, pp. 2688–2697, PMLR, 2018.

[27] H. Zhao, J. Gao, T. Lan, C. Sun, B. Sapp, B. Varadarajan, Y. Shen, Y. Shen, Y. Chai, C. Schmid, *et al.*, "Tnt: Target-driven trajectory prediction," *arXiv preprint arXiv:2008.08294*, 2020.

[28] M. Liang, B. Yang, R. Hu, Y. Chen, R. Liao, S. Feng, and R. Urtasun, "Learning Lane Graph Representations for Motion Forecasting," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12347 LNCS, pp. 541–556, 2020.

[29] W. Zeng, M. Liang, R. Liao, and R. Urtasun, "Lanercnn: Distributed representations for graph-centric motion forecasting," *arXiv preprint arXiv:2101.06653*, 2021.

[30] 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.

[31] J. Gao, C. Sun, H. Zhao, Y. Shen, D. Anguelov, C. Li, and C. Schmid, "Vectornet: Encoding hd maps and agent dynamics from vectorized representation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11522–11530, 2020.

[32] P. Bender, J. Ziegler, and C. Stiller, "Lanelets: Efficient map representation for autonomous driving," in *2014 IEEE Intelligent Vehicles Symposium Proceedings*, pp. 420–425, 2014.

[33] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

[34] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, 2019.

[35] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Konigshof, C. Stiller, A. de La Fortelle, *et al.*, "Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps," *arXiv preprint arXiv:1910.03088*, 2019.

[36] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8748–8757, 2019.

[37] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 261–268, 2009.

[38] A. Lerner, Y. Chrysanthou, and D. Lischinski, "Crowds by example," in *Computer graphics forum*, vol. 26, pp. 655–664, Wiley Online Library, 2007.

[39] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "Sophie: An attentive gan for predicting paths compliant to social and physical constraints," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1349–1358, 2019.

[40] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[41] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[42] R. Liaw, E. Liang, R. Nishihara, P. Moritz, J. E. Gonzalez, and I. Stoica, "Tune: A research platform for distributed model selection and training," *arXiv preprint arXiv:1807.05118*, 2018.

[43] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[45] M. Ye, T. Cao, and Q. Chen, "Tpcn: Temporal point cloud networks for motion forecasting," *arXiv preprint arXiv:2103.03067*, 2021.

[46] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso, "Transformer networks for trajectory forecasting," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 10335–10342, IEEE, 2021.

[47] C. Yu, X. Ma, J. Ren, H. Zhao, and S. Yi, "Spatio-temporal graph transformer networks for pedestrian trajectory prediction," in *European Conference on Computer Vision*, pp. 507–523, Springer, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In Sec. 6, we mention that our findings are limited to the particular model architecture and tasks studied in this work.
   (c) Did you discuss any potential negative societal impacts of your work? [No] We believe our work should not have any negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] See supplementary materials.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Because the code is proprietary.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See supplementary materials.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We do not estimate the amount of computation but provide the specification of the computational resources.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See supplementary materials.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] Yes, we mention how we obtained the data in the supplementary materials. Most datasets and packages we used are open-source. For the remaining INTERACTION dataset which is not open-source, we mention that we have been granted the access for research usage.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] We checked the format of the information provided by the datasets. There are no elements from the dataset that could potentially contain any personal information or offensive content.

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]