

# SceneGen: Learning to Generate Realistic Traffic Scenes

Shuhan Tan<sup>1,2\*</sup> Kelvin Wong<sup>1,3\*</sup> Shenlong Wang<sup>1,3</sup>

Sivabalan Manivasagam<sup>1,3</sup> Mengye Ren<sup>1,3</sup> Raquel Urtasun<sup>1,3</sup>

<sup>1</sup>Uber Advanced Technologies Group <sup>2</sup>Sun Yat-Sen University <sup>3</sup>University of Toronto

tanshh@mail2.sysu.edu.cn {kelvinwong, slwang, manivasagam, mren, urtasun}@cs.toronto.edu

## Abstract

We consider the problem of generating realistic traffic scenes automatically. Existing methods typically insert actors into the scene according to a set of hand-crafted heuristics and are limited in their ability to model the true complexity and diversity of real traffic scenes, thus inducing a content gap between synthesized traffic scenes versus real ones. As a result, existing simulators lack the fidelity necessary to train and test self-driving vehicles. To address this limitation, we present SceneGen—a neural autoregressive model of traffic scenes that eschews the need for rules and heuristics. In particular, given the ego-vehicle state and a high definition map of surrounding area, SceneGen inserts actors of various classes into the scene and synthesizes their sizes, orientations, and velocities. We demonstrate on two large-scale datasets SceneGen’s ability to faithfully model distributions of real traffic scenes. Moreover, we show that SceneGen coupled with sensor simulation can be used to train perception models that generalize to the real world.

## 1. Introduction

The ability to simulate realistic traffic scenarios is an important milestone on the path towards safe and scalable self-driving. It enables us to build rich virtual environments in which we can improve our self-driving vehicles (SDVs) and verify their safety and performance [9, 31, 32, 53]. This goal, however, is challenging to achieve. As a first step, most large-scale self-driving programs simulate pre-recorded scenarios captured in the real world [32] or employ teams of test engineers to design new scenarios [9, 31]. Although this approach can yield realistic simulations, it is ultimately not scalable. This motivates the search for a way to generate realistic traffic scenarios *automatically*.

More concretely, we are interested in generating the layout of actors in a traffic scene given the SDV’s current state and a high definition map (HD map) of the surround-

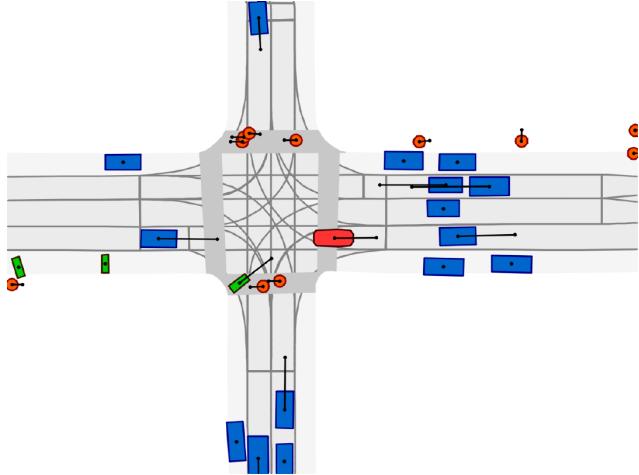


Figure 1: Given the SDV’s state and an HD map, SceneGen autoregressively inserts actors onto the map to compose a realistic traffic scene. The ego SDV is shown in red; vehicles in blue; pedestrians in orange; and bicyclists in green.

ing area. We call this task *traffic scene generation* (see Fig. 1). Here, each actor is parameterized by a class label, a bird’s eye view bounding box, and a velocity vector. Our lightweight scene parameterization is popular among existing self-driving simulation stacks and can be readily used in downstream modules; *e.g.*, to simulate LiDAR [9, 10, 32].

A popular approach to traffic scene generation is to use procedural models to insert actors into the scene according to a set of rules [55, 31, 9, 37]. These rules encode reasonable heuristics such as “pedestrians should stay on the sidewalk” or “vehicles should drive along lane centerlines”, and their parameters can be manually tuned to give reasonable results. Still, these simplistic heuristics cannot fully capture the complexity and diversity of real world traffic scenes, thus inducing a content gap between synthesized traffic scenes and real ones [26]. Moreover, this approach requires significant time and expertise to design good heuristics and tune their parameters.

To address these issues, recent methods use machine

\*Indicates equal contribution. Work done at Uber ATG.

learning techniques to automatically tune model parameters [52, 51, 24, 26, 8]. These methods improve the realism and scalability of traffic scene generation. However, they remain limited by their underlying hand-crafted heuristics and priors; *e.g.*, pre-defined scene grammars or assumptions about road topologies. As a result, they lack the capacity to model the true complexity and diversity of real traffic scenes and, by extension, the fidelity necessary to train and test SDVs in simulation. Alternatively, we can use a simple data-driven approach by sampling from map-specific empirical distributions [10]. But this cannot generalize to new maps and may yield scene-inconsistent samples.

In this paper, we propose SceneGen—a traffic scene generation model that eschews the need for hand-crafted rules and heuristics. Our approach is inspired by recent successes in deep generative modeling that have shown remarkable results in estimating distributions of a variety of data, without requiring complex rules and heuristics; *e.g.*, handwriting [18], images [49], text [39], *etc.* Specifically, SceneGen is a neural autoregressive model that, given the SDV’s current state and an HD map of the surrounding area, sequentially inserts actors into the scene—mimicking the process by which humans do this as well. As a result, we can sample realistic traffic scenes from SceneGen and compute the likelihood of existing ones as well.

We evaluate SceneGen on two large-scale self-driving datasets. The results show that SceneGen can better estimate the distribution over real traffic scenes than competing baselines and generate more realistic samples as well. Furthermore, we show that SceneGen coupled with sensor simulation can generate realistic labeled data to train perception models that generalize to the real world. With SceneGen, we take an important step towards developing SDVs safely and scalably through large-scale simulation. We hope our work here inspires more research along this direction so that one day this goal will become a reality.

## 2. Related Work

**Traffic simulation:** The study of traffic simulation can be traced back to at least the 1950s with Gerlough’s dissertation on simulating freeway traffic flow [16]. Since then, various traffic models have been used for simulation. Macroscopic models simulate entire populations of vehicles in the aggregate [30, 40] to study “macroscopic” properties of traffic flow, such as traffic density and average velocity. In contrast, microscopic models simulate the behavior of each individual vehicle over time by assuming a car-following model [36, 6, 34, 13, 17, 1, 44]. These models improve simulation fidelity considerably but at the cost of computational efficiency. Microscopic traffic models have been included in popular software packages such as SUMO [31], CORSIM [35], VISSIM [11], and MITSIM [55].

Recently, traffic simulation has found new applications

in testing and training the autonomy stack of SDVs. However, existing simulators do not satisfy the level of realism necessary to properly test SDVs [52]. For example, the CARLA simulator [9] spawns actors at pre-determined locations and uses a lane-following controller to simulate the vehicle behaviors over time. This approach is too simplistic and so it induces a sim2real content gap [26]. Therefore, in this paper, we study how to generate snapshots of traffic scenes that mimic the realism and diversity of real ones.

**Traffic scene generation:** While much of the research into microscopic traffic simulation have focused on modeling actors’ behaviors, an equally important yet underexplored problem is how to generate realistic snapshots of traffic scenes. These snapshots have many applications; *e.g.*, to initialize traffic simulations [52] or to generate labeled data for training perception models [26]. A popular approach is to procedurally insert actors into the scene according to a set of rules [55, 31, 9, 37]. These rules encode reasonable heuristics such as “pedestrians should stay on the sidewalk” and “vehicles should drive along lane center-lines”, and their parameters can be manually tuned to give reasonable results. For example, SUMO [31] inserts vehicles into lanes based on minimum headway requirements and initializes their speeds according to a Gaussian distribution [52]. Unfortunately, it is difficult to scale this approach to new environments since tuning these heuristics require significant time and expertise.

An alternative approach is to learn a probabilistic distribution over traffic scenes from which we can sample new scenes [52, 51, 24, 10, 14, 15, 57]. For example, Wheeler *et al.* [52] propose a Bayesian network to model a joint distribution over traffic scenes in straight multi-lane highways. This approach was extended to model inter-lane dependencies [51] and generalized to handle a four-way intersection [24]. These models are trained to mimic a real distribution over traffic scenes. However, they consider a limited set of road topologies only and assume that actors follow reference paths in the map. As a result, they are difficult to generalize to real urban scenes, where road topologies and actor behaviors are considerably more complex; *e.g.*, pedestrians do not follow reference paths in general.

Recent advances in deep learning have enabled a more flexible approach to learn a distribution over traffic scenes. In particular, MetaSim [26] augments the probabilistic scene graph of Prakash *et al.* [37] with a graph neural network. By modifying the scene graph’s node attributes, MetaSim reduces the content gap between synthesized images versus real ones, without manual tuning. MetaSim2 [8] extends this idea by learning to sample the scene graph as well. Unfortunately, these approaches are still limited by their hand-crafted scene grammar which, for example, constrains vehicles to lane centerlines. We aim to develop a

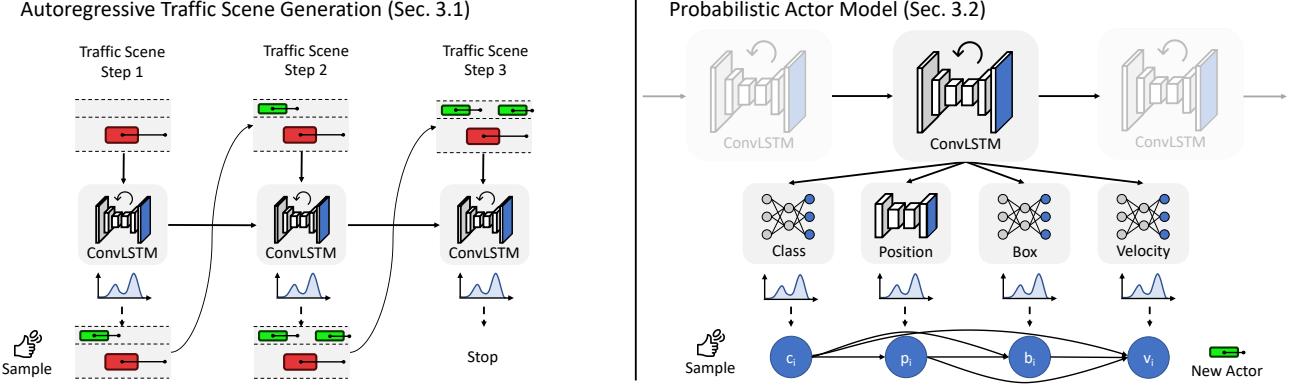


Figure 2: Overview of our approach. Given the ego SDV’s state and an HD map of the surrounding area, SceneGen generates a traffic scene by inserting actors one at a time (Sec. 3.1). We model each actor  $a_i \in \mathcal{A}$  probabilistically, as a product over distributions of its class  $c_i \in \mathbb{C}$ , position  $p_i \in \mathbb{R}^2$ , bounding box  $b_i \in \mathbb{B}$ , and velocity  $v_i \in \mathbb{R}^2$  (Sec. 3.2).

more general method that avoids requiring these heuristics.

**Autoregressive models:** Autoregressive models factorize a joint distribution over  $n$ -dimensions into a product of conditional distributions  $p(x) = \prod_{i=1}^n p(x_i|x_{<i})$ . Each conditional distribution is then approximated with a parameterized function [12, 2, 45, 46, 47]. Recently, neural autoregressive models have found tremendous success in modeling a variety of data, including handwriting [18], images [49], audio [48], text [39], sketches [20], graphs [29], 3D meshes [33], indoor scenes [50] and image scene layouts [25]. These models are particularly popular since they can factorize a complex joint distribution into a product of much simpler conditional distributions. Moreover, they generally admit a tractable likelihood, which can be used for likelihood-based training, uncovering interesting/outlier examples, *etc.* Inspired by these advances, we exploit autoregressive models for traffic scene generation as well.

### 3. Traffic Scene Generation

Our goal is to learn a distribution over traffic scenes from which we can sample new examples and evaluate the likelihood of existing ones. In particular, given the SDV  $a_0 \in \mathcal{A}$  and an HD map  $m \in \mathcal{M}$ , we aim to estimate the joint distribution over other actors in the scene  $\{a_1, \dots, a_n\} \subset \mathcal{A}$ ,

$$p(a_1, \dots, a_n | m, a_0) \quad (1)$$

The HD map  $m \in \mathcal{M}$  is a collection of polygons and polylines that provide semantic priors for a region of interest around the SDV; *e.g.*, lane boundaries, drivable areas, traffic light states. These priors provide important contextual information about the scene and allow us to generate actors that are consistent with the underlying road topology.

We parameterize each actor  $a_i \in \mathcal{A}$  with an eight-dimensional random variable containing its class label  $c_i \in$

$\mathbb{C}$ , its bird’s eye view location  $(x_i, y_i) \in \mathbb{R}^2$ , its bounding box  $b_i \in \mathbb{B}^1$ , and its velocity  $v_i \in \mathbb{R}^2$ . Each bounding box  $b_i \in \mathbb{B}$  is a 3-tuple consisting of the bounding box’s size  $(w_i, l_i) \in \mathbb{R}_{>0}^2$  and heading angle  $\theta_i \in [0, 2\pi]$ . In our experiments,  $\mathbb{C}$  consists of three classes: vehicles, pedestrians, and bicyclists. See Fig. 1 for an example.

Modeling Eq. 1 is a challenging task since the actors in a given scene are highly correlated among themselves and with the map, and the number of actors in the scene is random as well. We aim to model Eq. 1 such that our model is easy to sample from and the resulting samples reflect the complexity and diversity of real traffic scenes. Our approach is to autoregressively factorize Eq. 1 into a product of conditional distributions. This yields a natural generation process that sequentially inserts actors into the scene one at a time. See Fig. 2 for an overview of our approach.

In the following, we first describe our autoregressive factorization of Eq. 1 and how we model this with a recurrent neural network (Sec. 3.1). Then, in Sec. 3.2, we describe how SceneGen generates a new actor at each step of the generation process. Finally, in Sec. 3.3, we discuss how we train and sample from SceneGen.

#### 3.1. The Autoregressive Generation Process

Given the SDV  $a_0 \in \mathcal{A}$  and an HD map  $m \in \mathcal{M}$ , our goal is to estimate a conditional distribution over the actors in the scene  $\{a_1, \dots, a_n\} \subset \mathcal{A}$ . As we alluded to earlier, modeling this conditional distribution is challenging since the actors in a given scene are highly correlated among themselves and with the map, and the number of actors in the scene is random. Inspired by the recent successes of neural autoregressive models [18, 49, 39], we propose to autoregressively factorize  $p(a_1, \dots, a_n | m, a_0)$  into a product

<sup>1</sup>Pedestrians are not represented by bounding boxes. They are represented by a single point indicating their center of gravity.

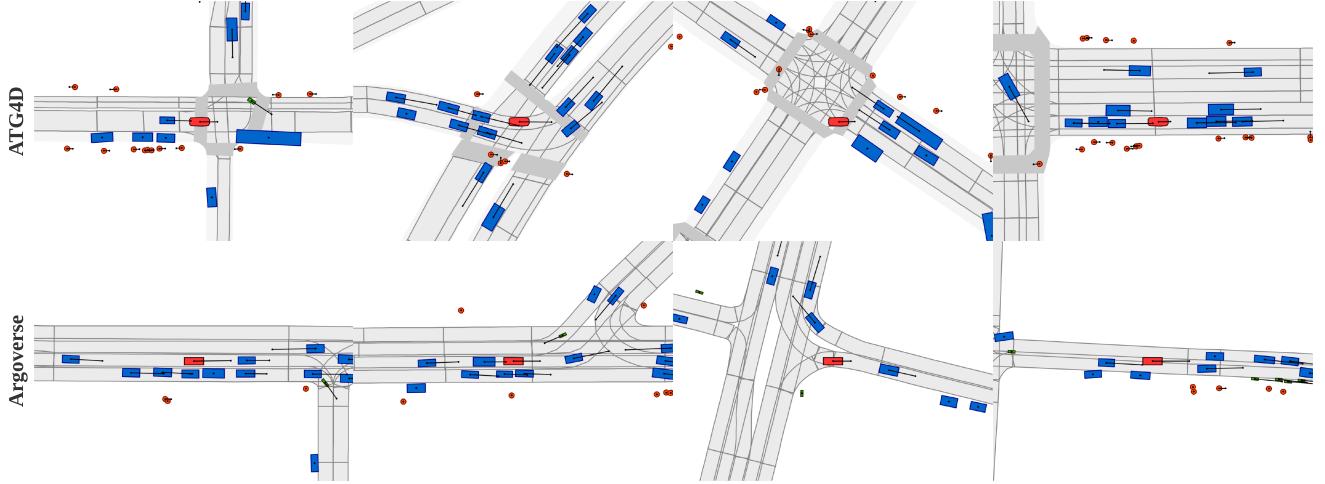


Figure 3: Traffic scenes generated by SceneGen conditioned on HD maps from ATG4D (top) and Argoverse (bottom).

of simpler conditional distributions. This factorization simplifies the task of modeling the complex joint distribution  $p(\mathbf{a}_1, \dots, \mathbf{a}_n | \mathbf{m}, \mathbf{a}_0)$  and results in a model with a tractable likelihood. Moreover, it yields a natural generation process that mimics how a human might perform this task as well.

In order to perform this factorization, we assume a fixed canonical ordering over the sequence of actors  $\mathbf{a}_1, \dots, \mathbf{a}_n$ ,

$$p(\mathbf{a}_1, \dots, \mathbf{a}_n | \mathbf{m}, \mathbf{a}_0) = p(\mathbf{a}_1 | \mathbf{m}, \mathbf{a}_0) \prod_{i=1}^n p(\mathbf{a}_i | \mathbf{a}_{<i}, \mathbf{m}, \mathbf{a}_0) \quad (2)$$

where  $\mathbf{a}_{<i} = \{\mathbf{a}_1, \dots, \mathbf{a}_{i-1}\}$  is the set of actors up to and including the  $i-1$ -th actor in canonical order. In our experiments, we choose a left-to-right, top-to-bottom order based on each actor’s position in bird’s eye view coordinates. We found that this intuitive ordering works well in practice.

Since the number of actors per scene is random, we introduce a stopping token  $\perp$  to indicate the end of our sequential generation process. In practice, we treat  $\perp$  as an auxillary actor that, when generated, ends the generation process. Therefore, for simplicity of notation, we assume that the last actor  $\mathbf{a}_n$  is always the stopping token  $\perp$ .

**Model architecture:** Our model uses a recurrent neural network to capture the long-range dependencies across our autoregressive generation process. The basis of our model is the ConvLSTM architecture [42]—an extension of the classic LSTM architecture [22] to spatial data—and the input to our model at the  $i$ -th generation step is a bird’s eye view multi-channel image encoding the SDV  $\mathbf{a}_0$ , the HD map  $\mathbf{m}$ , and the actors generated so far  $\{\mathbf{a}_1, \dots, \mathbf{a}_{i-1}\}$ .

For the  $i$ -th step of the generation process: Let  $\mathbf{x}^{(i)} \in \mathbb{R}^{C \times H \times W}$  denote the multi-channel image, where  $C$  is the

number of feature channels and  $H \times W$  is the size of the image grid. Given the previous hidden and cell states  $\mathbf{h}^{(i-1)}$  and  $\mathbf{c}^{(i-1)}$ , the new hidden and cell states are given by:

$$\mathbf{h}^{(i)}, \mathbf{c}^{(i)} = \text{ConvLSTM}(\mathbf{x}^{(i)}, \mathbf{h}^{(i-1)}, \mathbf{c}^{(i-1)}; \mathbf{w}) \quad (3)$$

$$\mathbf{f}^{(i)} = \text{CNN}_b(\mathbf{h}^{(i)}; \mathbf{w}) \quad (4)$$

where ConvLSTM is a two-layer ConvLSTM,  $\text{CNN}_b$  is a five-layer convolutional neural network (CNN) that extract backbone features, and  $\mathbf{w}$  are the neural network parameters. The features  $\mathbf{f}^{(i)}$  summarize the generated scene so far  $\mathbf{a}_{<i}$ ,  $\mathbf{a}_0$ , and  $\mathbf{m}$ , and we use  $\mathbf{f}^{(i)}$  to predict the conditional distribution  $p(\mathbf{a}_i | \mathbf{a}_{<i}, \mathbf{m}, \mathbf{a}_0)$ , which we describe next. See our appendix for details.

### 3.2. A Probabilistic Model of Actors

Having specified the generation process, we now turn our attention to modeling each actor probabilistically. As discussed earlier, each actor  $\mathbf{a}_i \in \mathcal{A}$  is parameterized by its class label  $c_i \in \mathbb{C}$ , location  $(x_i, y_i) \in \mathbb{R}^2$ , oriented bounding box  $\mathbf{b}_i \in \mathbb{B}$  and velocity  $\mathbf{v}_i \in \mathbb{R}^2$ . To capture the dependencies between these attributes, we factorize  $p(\mathbf{a}_i | \mathbf{a}_{<i}, \mathbf{m}, \mathbf{a}_0)$  as follows:

$$p(\mathbf{a}_i) = p(c_i)p(x_i, y_i | c_i)p(\mathbf{b}_i | c_i, x_i, y_i)p(\mathbf{v}_i | c_i, x_i, y_i, \mathbf{b}_i) \quad (5)$$

where we dropped the condition on  $\mathbf{a}_{<i}$ ,  $\mathbf{m}$ , and  $\mathbf{a}_0$  to simplify notation. Thus, the distribution over an actor’s location is conditional on its class; its bounding box is conditional on its class and location; and its velocity is conditional on its class, location, and bounding box. Note that if  $\mathbf{a}_i$  is the stopping token  $\perp$ , we do not model its location, bounding box, and velocity. Instead, we have  $p(\mathbf{a}_i) = p(c_i)$ , where  $c_i$  is the auxillary class  $c_\perp$ .

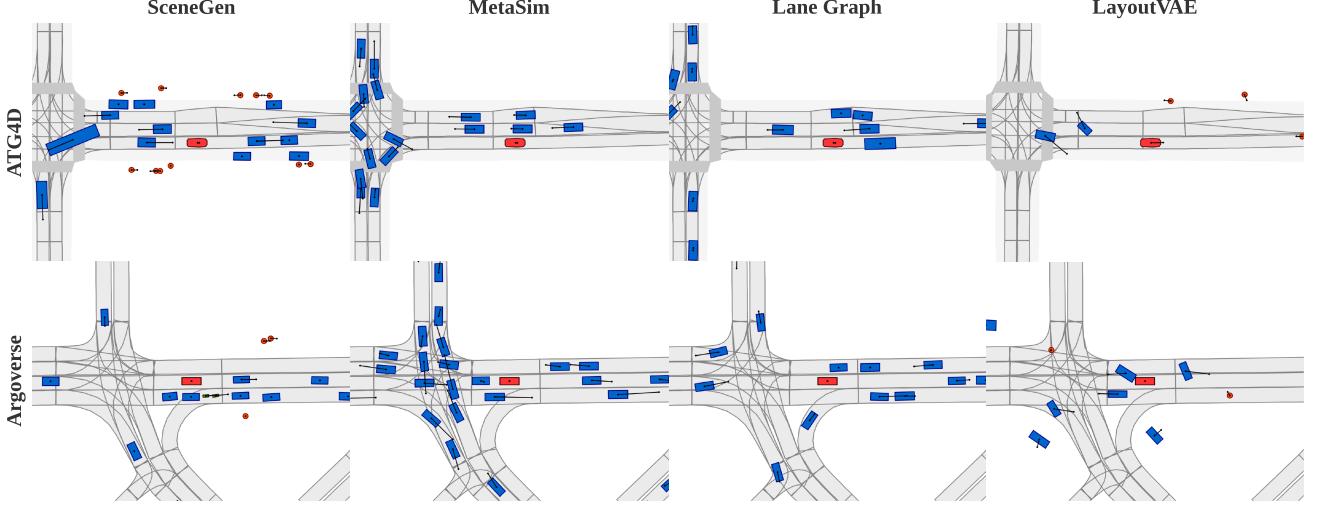


Figure 4: Qualitative comparison of traffic scenes generated by SceneGen and various baselines.

**Class:** To model a distribution over an actor’s class, we use a categorical distribution whose support is the set of classes  $\mathbb{C} \cup \{c_\perp\}$  and whose parameters  $\pi_c$  are predicted by a neural network:

$$\pi_c = \text{MLP}_c(\text{avg-pool}(\mathbf{f}^{(i)}); \mathbf{w}) \quad (6)$$

$$c_i \sim \text{Categorical}(\pi_c) \quad (7)$$

where  $\text{avg-pool}: \mathbb{R}^{C \times H \times W} \rightarrow \mathbb{R}^C$  is average pooling over the spatial dimensions and  $\text{MLP}_c$  is a three-layer multi-layer perceptron (MLP) with softmax activations.

**Location:** We apply uniform quantization to the actor’s position and model the quantized values using a categorical distribution. The support of this distribution is the set of  $H \times W$  quantized bins within our region of interest and its parameters  $\pi_{\text{loc}}$  are predicted by a class-specific CNN. This approach allows the model to express highly multi-modal distributions without making assumptions about the distribution’s shape [49]. To recover continuous values, we assume a uniform distribution within each quantization bin.

Let  $k$  denote an index into one of the  $H \times W$  quantized bins, and suppose  $[\mathbf{p}_k] \in \mathbb{R}^2$  (*resp.*,  $[\mathbf{p}_k] \in \mathbb{R}^2$ ) is the minimum (*resp.*, maximum) continuous coordinates in the  $k$ -th bin. We model  $p(x_i, y_i | c_i)$  as follows:

$$\pi_{\text{loc}} = \text{CNN}_{\text{loc}}(\mathbf{f}^{(i)}; c_i, \mathbf{w}) \quad (8)$$

$$k \sim \text{Categorical}(\pi_{\text{loc}}) \quad (9)$$

$$(x_i, y_i) \sim \text{Uniform}([\mathbf{p}_k], [\mathbf{p}_k]) \quad (10)$$

where  $\text{CNN}_{\text{loc}}(\cdot; c_i, \mathbf{w})$  is a CNN with softmax activations for the class  $c_i$ . During inference, we mask and re-normalize  $\pi_{\text{loc}}$  such that quantized bins with invalid posi-

tions according to our canonical ordering have zero probability mass. Note that we do not mask during training since this resulted in worse performance.

After sampling the actor’s location  $(x_i, y_i) \in \mathbb{R}^2$ , we extract a feature vector  $\mathbf{f}_{x_i, y_i}^{(i)} \in \mathbb{R}^C$  by spatially indexing into the  $k$ -th bin of  $\mathbf{f}^{(i)}$ . This feature vector captures local information at  $(x_i, y_i)$  and is used to subsequently predict the actor’s bounding box and velocity.

**Bounding box:** An actor’s bounding box  $\mathbf{b}_i \in \mathbb{B}$  consists of its width and height  $(w_i, l_i) \in \mathbb{R}_{>0}^2$  and its heading  $\theta_i \in [0, 2\pi]$ . We model the distributions over each of these independently. For an actor’s bounding box size, we use a mixture of  $K$  bivariate log-normal distributions:

$$[\pi_{\text{box}}, \mu_{\text{box}}, \Sigma_{\text{box}}] = \text{MLP}_{\text{box}}(\mathbf{f}_{x_i, y_i}^{(i)}; c_i, \mathbf{w}) \quad (11)$$

$$k \sim \text{Categorical}(\pi_{\text{box}}) \quad (12)$$

$$(w_i, l_i) \sim \text{LogNormal}(\mu_{\text{box}, k}, \Sigma_{\text{box}, k}) \quad (13)$$

where  $\pi_{\text{box}}$  are mixture weights, each  $\mu_{\text{box}, k} \in \mathbb{R}^2$  and  $\Sigma_{\text{box}, k} \in \mathbb{S}_+^2$  parameterize a component log-normal distribution, and  $\text{MLP}_{\text{box}}(\cdot; c_i, \mathbf{w})$  is a three-layer MLP for the class  $c_i$ . This parameterization allows our model to naturally capture the multi-modality of actor sizes in real world data; *e.g.*, the size of sedans versus trucks.

Similarly, we model an actor’s heading angle with a mixture of  $K$  Von-Mises distributions:

$$[\pi_\theta, \mu_\theta, \kappa_\theta] = \text{MLP}_\theta(\mathbf{f}_{x_i, y_i}^{(i)}; c_i, \mathbf{w}) \quad (14)$$

$$k \sim \text{Categorical}(\pi_\theta) \quad (15)$$

$$\theta_i \sim \text{VonMises}(\mu_{\theta, k}, \kappa_{\theta, k}) \quad (16)$$

where  $\pi_\theta$  are mixture weights, each  $\mu_{\theta, k} \in [0, 2\pi]$  and  $\kappa_{\theta, k} > 0$  parameterize a component Von-Mises distribu-

Method	ATG4D							Argoverse				
	NLL	Features	Class	Size	Speed	Heading	NLL	Features	Class	Size	Speed	Heading
Prob. Grammar	-	0.20	0.24	0.46	0.34	0.31	-	0.38	0.26	0.41	0.57	0.38
MetaSim	-	0.12	0.24	0.45	0.35	0.15	-	0.18	0.26	0.50	0.52	0.18
Procedural	-	0.38	0.24	0.17	0.34	<b>0.07</b>	-	0.58	0.26	0.23	0.59	0.17
Lane Graph	-	0.17	0.24	0.30	0.32	0.16	-	<b>0.11</b>	0.26	0.31	0.32	0.28
LayoutVAE	210.80	0.15	<b>0.12</b>	0.18	0.33	0.29	200.78	0.25	<b>0.11</b>	0.21	0.41	0.29
SceneGen	<b>59.86</b>	<b>0.11</b>	0.20	<b>0.06</b>	<b>0.27</b>	0.08	<b>67.11</b>	0.14	0.21	<b>0.17</b>	<b>0.17</b>	<b>0.21</b>

Table 1: Negative log-likelihood (NLL) and maximum mean discrepancy (MMD) results on ATG4D and Argoverse. NLL is reported in *nats*, averaged across all scenes in the test set. MMD is computed between distributions of features extracted by a motion forecasting model and various scene statistics (see main text for description). For all metrics, lower is better.

tion, and  $\text{MLP}_\theta(\cdot; c_i, \mathbf{w})$  is a three-layer MLP for the class  $c_i$ . The Von-Mises distribution is a close approximation of a normal distribution wrapped around the unit circle [38] and has the probability density function

$$p(\theta|\mu, \kappa) = \frac{e^{\kappa \cos(\theta - \mu)}}{2\pi I_0(\kappa)} \quad (17)$$

where  $I_0$  is the modified Bessel function of order 0. We use a mixture of Von-Mises distributions to capture the multimodality of headings in real world data; *e.g.*, a vehicle can go straight or turn at an intersection. To sample from a mixture of Von-Mises distributions, we sample a component  $k$  from a categorical distribution and then sample  $\theta$  from the Von-Mises distribution of the  $k$ -th component [3].

**Velocity:** We parameterize the actor’s velocity  $\mathbf{v}_i \in \mathbb{R}^2$  as  $\mathbf{v}_i = (s_i \cos \omega_i, s_i \sin \omega_i)$ , where  $s_i \in \mathbb{R}_{\geq 0}$  is its speed and  $\omega_i \in [0, 2\pi]$  is its direction. Note that this parameterization is not unique since  $\omega_i$  can take any value in  $[0, 2\pi]$  when  $\mathbf{v}_i = 0$ . Therefore, we model the actor’s velocity as a mixture model where one of the  $K \geq 2$  components corresponds to  $\mathbf{v}_i = 0$ . More concretely, we have

$$\boldsymbol{\pi}_v = \text{MLP}_v(\mathbf{f}_{x_i, y_i}^{(i)}; c_i, \mathbf{w}) \quad (18)$$

$$k \sim \text{Categorical}(\boldsymbol{\pi}_v) \quad (19)$$

where for  $k > 0$ , we have  $\mathbf{v}_i = (s_i \cos \omega_i, s_i \sin \omega_i)$ , with

$$[\mu_s, \sigma_s] = \text{MLP}_s(\mathbf{f}_{x_i, y_i}^{(i)}; c_i, \mathbf{w}) \quad (20)$$

$$[\mu_\omega, \kappa_\omega] = \text{MLP}_\omega(\mathbf{f}_{x_i, y_i}^{(i)}; c_i, \mathbf{w}) \quad (21)$$

$$s_i \sim \text{LogNormal}(\mu_{s,k}, \sigma_{s,k}) \quad (22)$$

$$\omega_i \sim \text{VonMises}(\mu_{\omega,k}, \kappa_{\omega,k}) \quad (23)$$

and for  $k = 0$ , we have  $\mathbf{v}_i = 0$ . As before, we use three-layer MLPs to predict the parameters of each distribution.

For vehicles and bicyclists, we parameterize  $\omega_i \in [0, 2\pi]$  as an offset relative to the actor’s heading  $\theta_i \in [0, 2\pi]$ . This is equivalent to parameterizing their velocities with a bicycle model [43], which we found improves sample quality.

### 3.3. Learning and Inference

**Sampling:** Pure sampling from deep autoregressive models can lead to degenerate examples due to their “unrealistic long tails” [23]. Therefore, we adopt a sampling strategy inspired by *nucleus sampling* [23]. Specifically, at each generation step, we sample from each of SceneGen’s output distributions  $M$  times and keep the most likely sample. We found this to help avoid degenerate traffic scenes while maintaining sample diversity. Furthermore, we reject vehicles and bicyclists whose bounding boxes collide with those of the actors sampled so far.

**Training:** We train our model to maximize the log-likelihood of real traffic scenes in our training dataset:

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} \sum_{i=1}^N \log p(\mathbf{a}_{i,1}, \dots, \mathbf{a}_{i,n} | \mathbf{m}_i, \mathbf{a}_{i,0}; \mathbf{w}) \quad (24)$$

where  $\mathbf{w}$  are the neural network parameters and  $N$  is the number of samples in our training set. In practice, we use the Adam optimizer [27] to minimize the average negative log-likelihood over mini-batches. We use teacher forcing and backpropagation-through-time to train through the generation process, up to a fixed window as memory allows.

## 4. Experiments

We evaluate SceneGen on two self-driving datasets: Argoverse [7] and ATG4D [54]. Our results show that SceneGen can generate more realistic traffic scenes than the competing methods (Sec. 4.3). We also demonstrate how SceneGen with sensor simulation can be used to train perception models that generalize to the real world (Sec. 4.4).

### 4.1. Datasets

**ATG4D:** ATG4D [54] is a large-scale dataset collected by a fleet of SDVs in cities across North America. It consists

# Mixtures	Scene	Vehicle	Pedestrian	Bicyclist
1	125.97	7.26	10.36	9.16
3	68.22	2.64	8.52	7.34
5	64.05	2.35	<b>8.27</b>	7.22
10	<b>59.86</b>	<b>1.94</b>	8.32	<b>6.90</b>

Table 2: Ablation of the number of mixture components in ATG4D. Scene NLL is averaged across scenes and NLL per class is the average NLL per actor of that class.

L	DA	C	TL	Scene	Veh.	Ped.	Bic.
				93.73	4.90	8.85	7.17
✓				63.33	2.12	8.69	7.10
✓	✓			<b>57.66</b>	<b>1.73</b>	8.40	6.84
✓	✓	✓		57.96	1.77	<b>8.32</b>	<b>6.61</b>
✓	✓	✓	✓	59.86	1.94	<b>8.32</b>	6.90

Table 3: Ablation of map in ATG4D (in NLL). **L** is lanes; **DA** drivable areas; **C** crosswalks; and **TL** traffic lights.

of 5500 25-seconds logs which we split into a training set of 5000 and an evaluation set of 500. Each log is subsampled at 10Hz to yield 250 traffic scenes, and each scene is annotated with bounding boxes for vehicles, pedestrians, and bicyclists. Each log also provides HD maps that encode lane boundaries, drivable areas, and crosswalks as polygons, and lane centerlines as polylines. Each lane segment is annotated with attributes such as its type (car *vs.* bike), turn direction, boundary colors, and traffic light state.

In our experiments, we subdivide the training set into two splits of 4000 and 1000 logs respectively. We use the first split to train the traffic scene generation models and the second split to train the perception models in Sec. 4.4.

**Argoverse:** Argoverse [7] consists of two datasets collected by a fleet of SDVs in Pittsburgh and Miami. We use the Argoverse 3D Tracking dataset which contains track annotations for 65 training logs and 24 validation logs. Each log is subsampled at 10Hz to yield 13,122 training scenes and 5015 validation scenes. As in ATG4D, Argoverse provides HD maps annotated with drivable areas and lane segment centerlines and their attributes; *e.g.*, turn direction. However, Argoverse does not provide crosswalk polygons, lane types, lane boundary colors, and traffic lights.

## 4.2. Experiment Setup

**Baselines:** Our first set of baselines is inspired by recent work on probabilistic scene grammars and graphs [37, 26, 8]. In particular, we design a probabilistic grammar of traffic scenes (**Prob. Grammar**) such that actors are randomly placed onto lane segments using a hand-crafted prior [37]. Sampling from this grammar yields a *scene graph*, and our

next baseline (**MetaSim**) uses a graph neural network to transform the attributes of each actor in the scene graph. Our implementation follows Kar *et al.* [26], except we use a training algorithm that is supervised with heuristically generated ground truth scene graphs.<sup>2</sup>

Our next set of baselines is inspired by methods that reason directly about the road topology of the scene [52, 51, 24, 32]. Given a *lane graph* of the scene, **Procedural** uses a set of rules to place actors such that they follow lane centerlines, maintain a minimum clearance to leading actors, *etc.* Each actor’s bounding box is sampled from a Gaussian KDE fitted to the training dataset [5] and velocities are set to satisfy speed limits and a time gap between successive actors on the lane graph. Similar to MetaSim, we also consider a learning-based version of Procedural that uses a lane graph neural network [28] to transform each actor’s position, bounding box, and velocity (**Lane Graph**).

Since the HD maps in ATG4D and Argoverse do not provide reference paths for pedestrians, the aforementioned baselines cannot generate pedestrians.<sup>3</sup> Therefore, we also compare against **LayoutVAE** [25]—a variational autoencoder for image layouts that we adapt for traffic scene generation. We modify LayoutVAE to condition on HD maps and output oriented bounding boxes and velocities for actors of every class. Please see our appendix for details.

**Metrics:** Our first metric measures the negative log-likelihood (**NLL**) of real traffic scenes from the evaluation distribution, measured in *nats*. NLL is a standard metric to compare generative models with tractable likelihoods. However, as many of our baselines do not have likelihoods, we compute a sample-based metric as well: maximum mean discrepancy (**MMD**) [19]. For two distributions  $p$  and  $q$ , MMD measures a distance between  $p$  and  $q$  as

$$\text{MMD}^2(p, q) = \mathbb{E}_{x, x' \sim p}[k(x, x')] + \mathbb{E}_{y, y' \sim q}[k(y, y')] - 2\mathbb{E}_{x \sim p, y \sim q}[k(x, y)] \quad (25)$$

for some kernel  $k$ . Following [56, 29], we compute MMD using Gaussian kernels with the total variation distance to compare distributions of scene statistics between generated and real traffic scenes. Our scene statistics measure the distribution of classes, bounding box sizes, speeds, and heading angles (relative to the SDV) in each scene. To peer into the global properties of the traffic scene, we also compute MMD in the feature space of a pre-trained motion forecasting model that takes a rasterized image of the scene as input [53]. This is akin to the popular IS [41], FID [21], and KID [4] metrics for evaluating generative models, except we use a feature extractor trained on traffic scenes. Please see our appendix for details.

<sup>2</sup>We were unable to train MetaSim using their unsupervised losses.

<sup>3</sup>In Argoverse, these baselines generate vehicles only since bike lanes are not given. This highlights the challenge of designing good heuristics.

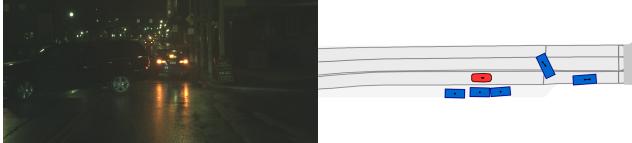


Figure 5: ATG4D scene with a traffic violation.

**Additional details:** Each traffic scene is a  $80m \times 80m$  region of interest centered on the ego SDV. By default, SceneGen uses  $K = 10$  mixture components and conditions on all available map elements for each dataset. We train SceneGen using the Adam optimizer [27] with a learning rate of  $1e-4$  and a batch size of 16, until convergence. When sampling each actor’s position, heading, and velocity, we sample  $M = 10$  times and keep the most likely sample.

### 4.3. Results

**Quantitative results:** Tab. 1 summarizes the NLL and MMD results for ATG4D and Argoverse. Overall, SceneGen achieves the best results across both datasets, demonstrating that it can better model real traffic scenes and synthesize realistic examples as well. Interestingly, all learning-based methods outperform the hand-tuned baselines with respect to MMD on deep features—a testament to the difficulty of designing good heuristics.

**Qualitative results:** Fig. 3 visualizes samples generated by SceneGen on ATG4D and Argoverse. Fig. 4 compares traffic scenes generated by SceneGen and various baselines. Although MetaSim and Lane Graph generate reasonable scenes, they are limited by their underlying heuristics; *e.g.*, actors follow lane centerlines. LayoutVAE generates a greater variety of actors; however, the model does not position actors on the map accurately, rendering the overall scene unrealistic. In contrast, SceneGen’s samples reflects the complexity of real traffic scenes much better. That said, SceneGen occasionally generates near-collision scenes that are plausible but unlikely; *e.g.*, Fig. 3 top-right.

**Ablation studies:** In Tab. 2, we sweep over the number of mixture components used to parameterize distributions of bounding boxes and velocities. We see that increasing the number of components consistently lowers NLL, reflecting the need to model the multi-modality of real traffic scenes. We also ablate the input map to SceneGen: starting from an unconditional model, we progressively add lanes, drivable areas, crosswalks, and traffic light states. From Tab. 3, we see that using more map elements generally improves NLL. Surprisingly, incorporating traffic lights slightly degrades performance, which we conjecture is due to infrequent traffic light observations in ATG4D.

Method	Vehicle		Pedestrian		Bicyclist	
	0.5	0.7	0.3	0.5	0.3	0.5
Prob. Gram.	81.1	66.6	-	-	11.2	10.6
MetaSim	76.3	63.3	-	-	8.2	7.5
Procedural	80.2	63.0	-	-	6.5	3.8
Lane Graph	82.9	71.7	-	-	7.6	6.9
LayoutVAE	85.9	76.3	49.3	41.8	18.4	16.4
SceneGen	<b>90.4</b>	<b>82.4</b>	<b>58.1</b>	<b>48.7</b>	<b>19.6</b>	<b>17.9</b>
Real Scenes	93.7	86.7	69.3	61.6	29.2	25.9

Table 4: Detection AP on real ATG4D scenes.



Figure 6: Outputs of detector trained with SceneGen scenes.

**Discovering interesting scenes:** We use SceneGen to find unlikely scenes in ATG4D by searching for scenes with the highest NLL, normalized by the number of actors. Fig. 5 shows an example of a traffic violation found via this procedure; the violating actor has an NLL of 21.28.

### 4.4. Sim2Real Evaluation

Our next experiment demonstrates that SceneGen coupled with sensor simulation can generate realistic labeled data for training perception models. For each method under evaluation, we generate 250,000 traffic scenes conditioned on the SDV and HD map in each frame of the 1000 held-out logs in ATG4D. Next, we use LiDARsim [32] to simulate the LiDAR point cloud corresponding to each scene. Finally, we train a 3D object detector [54] using the simulated LiDAR and evaluate its performance on real scenes and LiDAR in ATG4D.

From Tab. 4, we see that SceneGen’s traffic scenes exhibit the lowest sim2real gap. Here, Real Scenes is simulated LiDAR from ground truth placements. This reaffirms our claim that the underlying rules and priors used in MetaSim and Lane Graph induce a content gap. By eschewing these heuristics altogether, SceneGen learns to generate significantly more realistic traffic scenes. Intriguingly, LayoutVAE performs competitively despite struggling to position actors on the map. We conjecture that this is because LayoutVAE captures the diversity of actor classes, sizes, headings, *etc.* well. However, by accurately modeling actor positions as well, SceneGen further reduces the sim2real gap, as compared to ground truth traffic scenes.

## 5. Conclusion

We have presented SceneGen—a neural autoregressive model of traffic scenes from which we can sample new examples as well as evaluate the likelihood of existing ones. Unlike prior methods, SceneGen eschews the need for rules or heuristics, making it a more flexible and scalable approach for modeling the complexity and diversity of real world traffic scenes. As a result, SceneGen is able to generate realistic traffic scenes, thus taking an important step towards safe and scalable self-driving.

## References

- [1] M. Bando, K. Hasebe, A. Nakayama, A. Shibata, and Y. Sugiyama. Dynamical model of traffic congestion and numerical simulation. *Physical Review E*, 1995. [2](#)
- [2] Yoshua Bengio and Samy Bengio. Modeling high-dimensional discrete data with multi-layer neural networks. In *NeurIPS*, 1999. [3](#)
- [3] Donald Best and Nicholas Fisher. Efficient simulation of the von mises distribution. *Journal of the Royal Statistical Society. Series C. Applied Statistics*, 1979. [6](#)
- [4] Mikolaj Binkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD gans. In *ICLR*, 2018. [7](#)
- [5] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. 2006. [7](#)
- [6] Robert E. Chandler, Robert Herman, and Elliott W. Montroll. Traffic dynamics: Studies in car following. In *Operations Research*, 1958. [2](#)
- [7] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *CVPR*, 2019. [6](#), [7](#)
- [8] Jeevan Devarajan, Amlan Kar, and Sanja Fidler. Metasim2: Unsupervised learning of scene structure for synthetic data generation. 2020. [2](#), [7](#)
- [9] Alexey Dosovitskiy, Germán Ros, Felipe Codevilla, Antonio M. López, and Vladlen Koltun. CARLA: an open urban driving simulator. In *CoRL*, 2017. [1](#), [2](#)
- [10] Jin Fang, Dingfu Zhou, Feilong Yan, Tongtong Zhao, Feihu Zhang, Yu Ma, Liang Wang, and Ruigang Yang. Augmented lidar simulator for autonomous driving, 2019. [1](#), [2](#)
- [11] Martin Fellendorf. Vissim: A microscopic simulation tool to evaluate actuated signal control including bus priority. 1994. [2](#)
- [12] Brendan J. Frey, Geoffrey E. Hinton, and Peter Dayan. Does the wake-sleep algorithm produce good density estimators? In *NeurIPS*, 1995. [3](#)
- [13] Denos C. Gazis, Robert Herman, and Richard W. Rothery. Nonlinear follow-the-leader models of traffic flow. In *Operations Research*, 1961. [2](#)
- [14] Andreas Geiger, Martin Lauer, and Raquel Urtasun. A generative model for 3d urban scene understanding from movable platforms. In *CVPR*, 2011. [2](#)
- [15] Andreas Geiger, Christian Wojek, and Raquel Urtasun. Joint 3d estimation of objects and scene layout. In *NeurIPS*, 2011. [2](#)
- [16] Daniel Gerlough. *Simulation of Freeway Traffic on a General-purpose Discrete Variable Computer*. 1955. [2](#)
- [17] Peter Gipps. Computer program multsim for simulating output from vehicle detectors on a multi-lane signal-controlled road. 1976. [2](#)
- [18] Alex Graves. Generating sequences with recurrent neural networks. *CoRR*, 2013. [2](#), [3](#)
- [19] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *JMLR*, 2012. [7](#)
- [20] David Ha and Douglas Eck. A neural representation of sketch drawings. In *ICLR*, 2018. [3](#)
- [21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. [7](#)
- [22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997. [4](#)
- [23] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, 2020. [6](#)
- [24] Stefan Jesenski, Jan Erik Stellet, Florian A. Schiegg, and J. Marius Zöllner. Generation of scenes in intersections for the validation of highly automated driving functions. In *IV*, 2019. [2](#), [7](#)
- [25] Akash Abdu Jyothi, Thibaut Durand, Jiawei He, Leonid Sigal, and Greg Mori. Layoutvae: Stochastic scene layout generation from a label set. In *ICCV*, 2019. [3](#), [7](#)
- [26] Amlan Kar, Aayush Prakash, Ming-Yu Liu, Eric Cameracci, Justin Yuan, Matt Rusiniak, David Acuna, Antonio Torralba, and Sanja Fidler. Meta-sim: Learning to generate synthetic datasets. In *ICCV*, 2019. [1](#), [2](#), [7](#)
- [27] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. [6](#), [8](#)
- [28] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *ECCV*, 2020. [7](#)
- [29] Renjie Liao, Yujia Li, Yang Song, Shenlong Wang, William L. Hamilton, David Duvenaud, Raquel Urtasun, and Richard S. Zemel. Efficient graph generation with graph recurrent attention networks. In *NeurIPS*, 2019. [3](#), [7](#)
- [30] Michael James Lighthill and Gerald Beresford Whitham. On kinematic waves. II. A theory of traffic flow on long crowded roads. In *Royal Society of London. Series A, Mathematical and Physical Sciences*, 1955. [2](#)
- [31] Pablo Álvarez López, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücke, Johannes Rummel, Peter Wagner, and Eva-Maria Wießner. Microscopic traffic simulation using SUMO. In *ITSC*, 2018. [1](#), [2](#)
- [32] Sivabalan Manivasagam, Shenlong Wang, Kelvin Wong, Wenyuan Zeng, Mikita Sazanovich, Shuhan Tan, Bin Yang, Wei-Chiu Ma, and Raquel Urtasun. Lidarsim: Realistic lidar simulation by leveraging the real world. In *CVPR*, 2020. [1](#), [7](#), [8](#)

- [33] Charlie Nash, Yaroslav Ganin, S. M. Ali Eslami, and Peter W. Battaglia. Polygen: An autoregressive generative model of 3d meshes. *ICML*, 2020. 3
- [34] G. F. Newell. Nonlinear effects in the dynamics of car following. *Operations Research*, 1961. 2
- [35] Larry Owen, Yunlong Zhang, Lei Rao, and Gene Mchale. Traffic flow simulation using corsim. *Winter Simulation Conference*, 2001. 2
- [36] Louis A. Pipes. An operational analysis of traffic dynamics. In *Journal of Applied Physics*, 1953. 2
- [37] Aayush Prakash, Shaad Boochoon, Mark Brophy, David Acuna, Eric Cameracci, Gavriel State, Omer Shapira, and Stan Birchfield. Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In *ICRA*, 2019. 1, 2, 7
- [38] Sergey Prokudin, Peter V. Gehler, and Sebastian Nowozin. Deep directional statistics: Pose estimation with uncertainty quantification. In *ECCV*, 2018. 6
- [39] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018. 2, 3
- [40] Paul Richards. Shock waves on the highway. In *Operations Research*, 1956. 2
- [41] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 7
- [42] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, 2015. 4
- [43] Saied Taheri. *An investigation and design of slip control braking systems integrated with four wheel steering*. 1990. 6
- [44] Martin Treiber, Ansgar Hennecke, and Dirk Helbing. Congested traffic states in empirical observations and microscopic simulations. In *Automatisierungstechnik*, 2000. 2
- [45] Benigno Uria, Iain Murray, and Hugo Larochelle. NADE: the real-valued neural autoregressive density-estimator. *CoRR*, 2013. 3
- [46] Benigno Uria, Iain Murray, and Hugo Larochelle. RNADE: the real-valued neural autoregressive density-estimator. In *NeurIPS*, 2013. 3
- [47] Benigno Uria, Iain Murray, and Hugo Larochelle. A deep and tractable density estimator. In *ICML*, 2014. 3
- [48] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *ISCA*, 2016. 3
- [49] Aäron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. In *ICML*, 2016. 2, 3, 5
- [50] Kai Wang, Manolis Savva, Angel X. Chang, and Daniel Ritchie. Deep convolutional priors for indoor scene synthesis. *TOG*, 2018. 3
- [51] Tim Allan Wheeler and Mykel J. Kochenderfer. Factor graph scene distributions for automotive safety analysis. In *ITSC*, 2016. 2, 7
- [52] Tim Allan Wheeler, Mykel J. Kochenderfer, and Philipp Robbel. Initial scene configurations for highway traffic propagation. In *ITSC*, 2015. 2, 7
- [53] Kelvin Wong, Qiang Zhang, Ming Liang, Bin Yang, Renjie Liao, Abbas Sadat, and Raquel Urtasun. Testing the safety of self-driving vehicles by simulating perception and prediction. *ECCV*, 2020. 1, 7
- [54] Bin Yang, Ming Liang, and Raquel Urtasun. HDNET: exploiting HD maps for 3d object detection. In *CoRL*, 2018. 6, 8
- [55] Qi Yang and Haris N. Koutsopoulos. A microscopic traffic simulator for evaluation of dynamic traffic management systems. *Transportation Research Part C: Emerging Technologies*, 1996. 1, 2
- [56] Jiaxuan You, Rex Ying, Xiang Ren, William L. Hamilton, and Jure Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models. In *ICML*, 2018. 7
- [57] Hongyi Zhang, Andreas Geiger, and Raquel Urtasun. Understanding high-level semantics by modeling traffic patterns. In *ICCV*, 2013. 2

## Supplementary Materials

### SceneGen: Learning to Generate Realistic Traffic Scenes

Shuhan Tan<sup>1,2\*</sup> Kelvin Wong<sup>1,3\*</sup> Shenlong Wang<sup>1,3</sup>

Sivabalan Manivasagam<sup>1,3</sup> Mengye Ren<sup>1,3</sup> Raquel Urtasun<sup>1,3</sup>

<sup>1</sup>Uber Advanced Technologies Group <sup>2</sup>Sun Yat-Sen University <sup>3</sup>University of Toronto

tanshh@mail2.sysu.edu.cn {kelvinwong, slwang, manivasagam, mren, urtasun}@cs.toronto.edu

### Abstract

In our supplementary materials, we detail SceneGen’s model architecture and training procedure (Sec. 1). Additionally, we provide additional experiment details in Sec. 2 and additional experiment results in Sec. 3. In Sec. 4, we exhibit an extensive array of qualitative results that demonstrate the realism and diversity of the traffic scenes generated by SceneGen.

## 1. Additional Model Details

### 1.1. Input Representation

At each step of the generation process, SceneGen takes a bird’s eye view multi-channel image encoding the HD map  $\mathbf{m}$ , the SDV  $\mathbf{a}_0$ , and the actors generated so far  $\{\mathbf{a}_1, \dots, \mathbf{a}_{i-1}\}$ . The image encompasses an  $80m \times 80m$  region of interest centered on the SDV  $\mathbf{a}_0$  and has a resolution of  $0.25m$  per pixel, yielding a  $320 \times 320$  image. The HD map is rasterized into a multi-channel image describing all available map elements in each dataset. For ATG4D, our multi-channel image consists of: lane polygons (straight vehicle lanes, dedicated right vehicle lanes, dedicated left vehicle lanes, dedicated bus lanes, and dedicated bike lanes); lane centerlines and dividers (allowed to cross, forbidden to cross, and maybe allowed to cross); lane segments (straight vehicle lanes, dedicated right vehicle lanes, and dedicated left vehicle lanes); drivable area and road polygons; and crosswalk polygons. In addition, we also encode each lane segment’s traffic light state (green, yellow, red, flashing yellow, flashing red, and unknown), speed limit, and orientation as filled lane polygons. Note that orientation angles are encoded in their Biternion representations  $\theta = (\cos \theta, \sin \theta)$  [16]. In aggregate, this yields a 24-channel image.

Argoverse provides a more limited set of map elements. Here, our multi-channel image consists of: lane polygons; lane centerlines (all lanes, left turn lanes, right turn lanes, intersection lanes, and traffic-controlled lanes); lane orientations (in Biternion representation); and drivable area polygons. In aggregate, this yields a 9-channel image.

To encode the actors  $\mathbf{a}_0, \mathbf{a}_1, \dots$ , we rasterize their bounding boxes onto a collection of binary occupancy images [1], one for each class; *i.e.*, SDV, vehicles, pedestrians, and bicyclists. Furthermore, we encode their headings and velocities by rasterizing their bounding boxes onto a five-channel image, filled with their respective speed, direction, and heading. As before, direction and heading angles are encoded in their Biternion representations. See Fig. 1 for an example.

### 1.2. Model Architecture

The basis of our model is the ConvLSTM architecture [19]. Let  $\mathbf{x}^{(i)} \in \mathbb{R}^{C \times H \times W}$  denote the input multi-channel image at the  $i$ -th step of the generation process. Given the previous hidden and cell states  $\mathbf{h}^{(i-1)}$  and  $\mathbf{c}^{(i-1)}$ , the new hidden states  $\mathbf{h}^{(i)}$ , cell states  $\mathbf{c}^{(i)}$ , and backbone features  $\mathbf{f}^{(i)}$  are given by:

$$\mathbf{h}^{(i)}, \mathbf{c}^{(i)} = \text{ConvLSTM}(\mathbf{x}^{(i)}, \mathbf{h}^{(i-1)}, \mathbf{c}^{(i-1)}; \mathbf{w}) \quad (1)$$

$$\mathbf{f}^{(i)} = \text{CNN}_b(\mathbf{h}^{(i)}; \mathbf{w}) \quad (2)$$

---

\*Indicates equal contribution. Work done at Uber ATG.

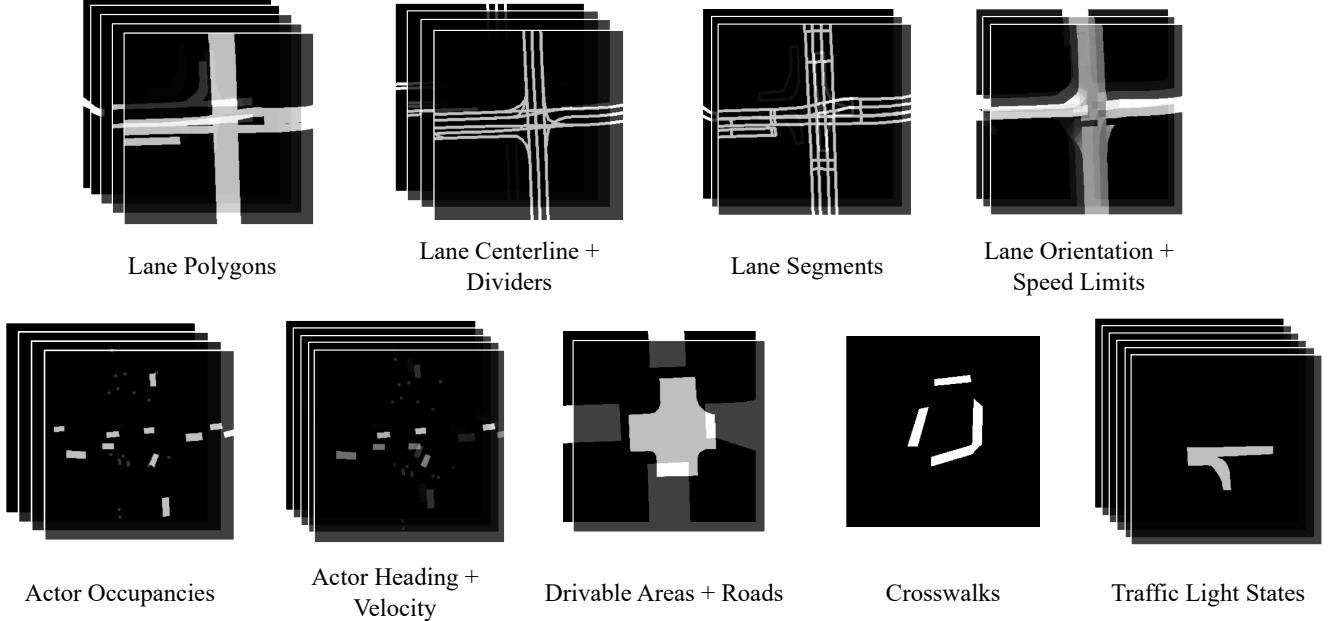


Figure 1: The input multi-channel image to SceneGen for ATG4D.

Here, ConvLSTM is a two-layer ConvLSTM with  $5 \times 5$  convolution kernels and 32 hidden channels, and  $\text{CNN}_b$  is a five-layer convolutional neural network (CNN) with 32 feature channels per layer. Each convolution layer consists of a  $3 \times 3$  convolution kernel, Group Normalization [23], and ReLU activations. The backbone features  $f^{(i)}$  summarize the generated scene so far and are given as input to the subsequent actor modules, which we detail next.

**Class:** We predict the class categorical distribution parameters  $\pi_c \in \Delta^{|\mathbb{C}|}$  as follows<sup>1</sup>:

$$\pi_c = \text{MLP}_c(\text{avg-pool}(f^{(i)}); \mathbf{w}) \quad (3)$$

where  $\text{avg-pool}: \mathbb{R}^{C \times H \times W} \rightarrow \mathbb{R}^C$  is average pooling over the spatial dimensions and  $\text{MLP}_c$  is a three-layer multi-layer perceptron (MLP) with 32 feature channels per hidden layer, ReLU activations, and softmax outputs.

**Location:** We apply uniform quantization to each actor’s position and model the quantized values with a categorical distribution. Our quantization resolution is  $0.25m$ , which we found sufficient to generate realistic traffic scenes while balancing memory efficiency. To predict the parameters  $\pi_{\text{loc}} \in \Delta^{H \times W - 1}$ , we use a three-layer CNN with 32 feature channels per hidden layer. Each hidden convolution layer consists of a  $3 \times 3$  convolution kernel, Group Normalization [23], and ReLU activations. The output convolution layer uses a  $1 \times 1$  kernel with softmax activations. Note that we use separate CNN weights for each class in  $\mathbb{C}$ ; *i.e.*, vehicles, pedestrians, and bicyclists.

**Bounding box:** An actor’s bounding box  $\mathbf{b}_i \in \mathbb{B}$  consists of its width and height  $(w_i, l_i) \in \mathbb{R}_{>0}^2$  and its heading  $\theta_i \in [0, 2\pi]$ . We model the distribution over bounding box sizes with a mixture of  $K$  bivariate log-normal distributions whose parameters are predicted by a three-layer MLP (with the same architecture as described earlier):

$$[\pi_{\text{box}}, \mu_{\text{box}}, \Sigma_{\text{box}}] = \text{MLP}_{\text{box}}(f_{x_i, y_i}^{(i)}; c_i, \mathbf{w}) \quad (4)$$

where  $\pi_{\text{box}} \in \Delta^{K-1}$  are mixture weights and each  $\mu_{\text{box}, k} \in \mathbb{R}^2$  and  $\Sigma_{\text{box}, k} \in \mathbb{S}_+^2$  parameterize a component log-normal distribution. To enforce the constraint that each  $\Sigma \in \mathbb{S}_+^2$ ,  $\text{MLP}_{\text{box}}$  predicts a variance term  $\sigma^2 \in \mathbb{R}_{>0}^2$  (in log-scale) and a

<sup>1</sup>We use  $\Delta^n = \{(x_0, x_1, \dots, x_n) \in \mathbb{R}^{n+1} \mid \sum_i x_i = 1 \text{ and } x_i \geq 0 \text{ for all } i\}$  to denote the  $n$ -simplex.

correlation term  $\rho \in [-1, 1]$  (using  $\tanh$ ) such that:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \in \mathbb{S}_+^2 \quad (5)$$

Similarly, we model the distribution over heading angles with a mixture of  $K$  Von-Mises distributions whose parameters are predicted by another three-layer MLP:

$$[\pi_\theta, \mu_\theta, \kappa_\theta] = \text{MLP}_\theta(\mathbf{f}_{x_i, y_i}^{(i)}; c_i, \mathbf{w}) \quad (6)$$

where  $\pi_\theta \in \Delta^{K-1}$  are mixture weights and each  $\mu_{\theta,k} \in [0, 2\pi)$  and  $\kappa_{\theta,k} > 0$  parameterize a component Von-Mises distribution. Following Prokudin *et al.* [16], we parameterize each  $\mu$  with its Baternion representation  $\mu = (\cos \mu, \sin \mu)$  and each  $\kappa$  is predicted in log-scale. Note that we use separate MLP weights for each class in  $\mathbb{C}$  whose actors are represented by bounding boxes; *i.e.*, vehicles and bicyclists. Pedestrians are represented by their center of gravity only (*i.e.*, location).

**Velocity:** Each of  $\text{MLP}_v$ ,  $\text{MLP}_s$ , and  $\text{MLP}_\omega$  is a three-layer MLP with the same architecture as described above. We parameterize the mixture of  $K$  Von-Mises distributions for directions  $\omega$  just as we parameterize the distribution of headings. As before, we use separate MLP weights for each class in  $\mathbb{C}$ .

### 1.3. Training Details

We train our model to maximize the log-likelihood of real traffic scenes in our training dataset:

$$\mathbf{w}^* = \arg \max_{\mathbf{w}} \sum_{i=1}^N \log p(\mathbf{a}_{i,1}, \dots, \mathbf{a}_{i,n} | \mathbf{m}_i, \mathbf{a}_{i,0}; \mathbf{w}) \quad (7)$$

where  $\mathbf{w}$  are the neural network parameters and  $N$  is the number of samples in our training set. We use teacher forcing and backpropagation-through-time to train through the generation process, up to a fixed window as memory allows. On a Nvidia Quadro RTX 5000 with 16GB of GPU memory, we train through 25 generation steps with batch size of 1 per GPU. We use PyTorch [14] and Horovod [18] to distribute the training process over 16 GPUs with a total batch size of 16. During training, we also randomly rotate each traffic scene with  $\theta \in [0, 2\pi)$ .

Note that each summand  $\log p(\mathbf{a}_1, \dots, \mathbf{a}_n | \mathbf{m}; \mathbf{w})$  can be decomposed into a sum of the log-likelihoods for each actors; namely, we have

$$\log p(\mathbf{a}_i | \xi_i) = \underbrace{\log p(c_i | \xi_i)}_{\text{class}} + \underbrace{\log p(x_i, y_i | c_i, \xi_i)}_{\text{location}} + \underbrace{\log p(\mathbf{b}_i | c_i, x_i, y_i, \xi_i)}_{\text{bounding box}} + \underbrace{\log p(\mathbf{v}_i | c_i, x_i, y_i, \mathbf{b}_i, \xi_i)}_{\text{velocity}} \quad (8)$$

where  $\xi_i$  encapsulates the conditions on  $\mathbf{a}_{<i}$ ,  $\mathbf{m}$ , and  $\mathbf{a}_0$ , to simplify notation. Therefore, the first summand  $\log p(c_i | \xi_i)$  is the (negative) cross-entropy loss between the predicted parameters  $\pi_c$  and the ground truth class  $c_i \in \mathbb{C} \cup \{\perp\}$ . We describe the remaining summands in detail next.

**Location:** The second summand  $\log p(x_i, y_i | c_i, \xi_i)$  measures the log-likelihood the actor's location  $(x_i, y_i) \in \mathbb{R}^2$ . As discussed earlier, we uniformly quantize each actor's location and parameterize it with a categorical distribution. Therefore,  $\log p(x_i, y_i | c_i, \xi_i)$  is the (negative) cross-entropy loss between the predicted parameters  $\pi_{\text{loc}}$  and the actor's ground truth quantized location. To address the significant imbalance of positive versus negative locations here, we use online negative hard mining. Specifically, we normalize  $\pi_{\text{loc}}$  over the hardest 10,000 locations only (including the positive location), and compute  $\log p(x_i, y_i | c_i, \xi_i)$  based this restricted categorical distribution instead.

**Bounding box:** The third summand  $\log p(\mathbf{b}_i | c_i, x_i, y_i, \xi_i)$  is a sum of the log-likelihoods of the actor's bounding box size  $(w_i, l_i) \in \mathbb{R}^2$  and heading  $\theta_i \in [0, 2\pi)$ :

$$\log p(\mathbf{b}_i | c_i, x_i, y_i, \xi_i) = \log p(w_i, l_i | c_i, x_i, y_i, \xi_i) + \log p(\theta_i | c_i, x_i, y_i, \xi_i) \quad (9)$$

Since we model bounding box size with a mixture of  $K$  bivariate log-normal distributions, we have

$$\log p(w_i, l_i | c_i, x_i, y_i, \xi_i) = \log \sum_{k=1}^K \pi_k \frac{1}{2\pi\sigma_{k,1}\sigma_{k,2}\sqrt{1-\rho_k^2}} e^{-\frac{1}{2(1-\rho_k^2)} \left[ \left(\frac{\log w_i - \mu_{k,1}}{\sigma_{k,1}}\right)^2 + \left(\frac{\log l_i - \mu_{k,2}}{\sigma_{k,2}}\right)^2 + 2\rho_k \left(\frac{\log w_i - \mu_{k,1}}{\sigma_{k,1}}\right) \left(\frac{\log l_i - \mu_{k,2}}{\sigma_{k,2}}\right) \right]} \quad (10)$$

where  $\pi \in \Delta^{K-1}$  are mixture weights and each  $\mu_k \in \mathbb{R}^2$ ,  $\sigma_k \in \mathbb{R}_{>0}^2$ , and  $\rho_k \in [-1, 1]$  parameterize a component bivariate log-normal distribution.

Similarly, since we model heading angles with a mixture of  $K$  Von-Mises distributions, we have

$$\log p(\theta_i | c_i, x_i, y_i, \xi_i) = \log \sum_{k=1}^K \pi_k \frac{e^{\kappa_k \cos(\theta_i - \mu_k)}}{2\pi I_0(\kappa_k)} \quad (11)$$

where  $\pi \in \Delta^{K-1}$  are mixture weights and each  $\mu_k \in [0, 2\pi)$  and  $\kappa_k > 0$  parameterize a component Von-Mises distribution.

**Velocity:** The fourth summand  $\log p(\mathbf{v}_i | c_i, x_i, y_i, \mathbf{b}_i, \xi_i)$  is the log-likelihood of the actor's velocity  $\mathbf{v}_i \in \mathbb{R}^2$ , which we parameterize as  $\mathbf{v}_i = (s_i \cos \omega_i, s_i \sin \omega_i)$  where  $s_i \in \mathbb{R}_{\geq 0}$  is its speed and  $\omega_i \in [0, 2\pi)$  is its direction. Recall that we model the distribution over an actor's velocity as a mixture model where one of the  $K \geq 2$  components corresponds to  $\mathbf{v}_i = 0$ . Therefore, for  $\mathbf{v}_i = 0$ , we have

$$\log p(\mathbf{v}_i | c_i, x_i, y_i, \mathbf{b}_i, \xi_i) = \log \pi_1 \quad (12)$$

and for  $\mathbf{v}_i > 0$ , we have

$$\log p(\mathbf{v}_i | c_i, x_i, y_i, \mathbf{b}_i, \xi_i) = \log \sum_{k=2}^K \pi_k \underbrace{\frac{1}{\sigma_{s,k}\sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{\log s_i - \mu_{s,k}}{\sigma_{s,k}}\right)^2}}_{\text{speed}} \times \underbrace{\frac{e^{\kappa_{\omega,k} \cos(\omega_i - \mu_{\omega,k})}}{2\pi I_0(\kappa_{\omega,k})}}_{\text{direction}} \quad (13)$$

where  $\pi \in \Delta^{K-1}$  are mixture weights, each  $\mu_{s,k} \in \mathbb{R}$  and  $\sigma_{s,k} > 0$  parameterize a component log-normal distribution for speed  $s_i$ , and each  $\mu_{\omega,k} \in [0, 2\pi)$  and  $\kappa_{\omega,k} > 0$  parameterize a component Von-Mises distribution for direction  $\omega_i$ .

## 2. Additional Experiment Details

### 2.1. Baselines

**Prob. Grammar:** Our Prob. Grammar baseline is inspired by recent work on probabilistic scene grammars [15, 9, 3]. Here, traffic scenes are composed by placing actors onto lane segments in the HD map, and initializing their classes, sizes, headings, velocities according to a hand-crafted prior. In our experiments, we use the following scene grammar:

$$\langle \text{Scene} \rangle \rightarrow \langle \text{Lanes} \rangle \quad (14)$$

$$\langle \text{Lanes} \rangle \rightarrow \langle \text{Lane} \rangle \langle \text{Lanes} \rangle | \epsilon \quad (15)$$

$$\langle \text{Lane} \rangle \rightarrow \langle \text{Actors} \rangle \quad (16)$$

$$\langle \text{Actors} \rangle \rightarrow \langle \text{Actor} \rangle \langle \text{Actors} \rangle | \epsilon \quad (17)$$

where Actor and  $\epsilon$  are terminal symbols. Sampling from this scene grammar yields a *scene graph*, which defines the *scene structure*—where lane segments are and which actors are positioned on top of them—and *scene parameters*—the attributes of each lane segment and actor. In our setting, we are given the lane nodes (and the SDV actor's node) of the scene graph as a condition, and our goal is to insert/modify the actor nodes.

Drawing inspiration from MetaSim's probabilistic scene grammar [9], we first uniformly sample the maximum number of actors per lane segment and then place them along the lane centerline, with a random clearance between successive actors drawn from the exponential distribution. The class of each actor is determined by the lane segment under consideration (*i.e.*, car lane *vs.* bike lane); its lateral offset from the lane centerline is given by uniform noise; its bounding box size is sampled from a uniform distribution; its heading and the direction of its velocity is given by the direction of the lane segment plus some uniform noise; and its speed is the minimum of a sample from a uniform distribution and the lane segment's speed limit. The parameters of every distribution are tuned by hand.

Method	ATG4D			Argoverse		
	Size	Speed	Heading	Size	Speed	Heading
Prob. Grammar	0.49	0.42	0.30	0.41	0.57	0.38
MetaSim	0.49	0.33	0.14	0.50	0.53	<b>0.18</b>
Procedural	0.15	0.41	<b>0.07</b>	0.23	0.59	<b>0.17</b>
Lane Graph	0.33	0.28	0.16	0.31	0.34	0.38
LayoutVAE	0.16	0.40	0.29	0.21	0.46	0.29
SceneGen	<b>0.06</b>	<b>0.19</b>	0.08	<b>0.15</b>	<b>0.20</b>	0.22

Table 1: Vehicle-only maximum mean discrepancy (MMD) results on ATG4D and Argoverse.

**MetaSim:** Our next baseline (MetaSim) uses a graph neural network (GNN) to transform the attributes of each actor node in the given scene graph. We use the implementation of Kar *et al.* [9] for this purpose. Specifically, given a scene graph drawn from Prob. Grammar, MetaSim deterministically modifies each actor’s distance along its lane centerline, lateral offset, bounding box size, heading, and velocity. The inputs to MetaSim is a scene graph where each node’s features are its attributes (normalized between 0 and 1 based on their respective minimum/maximum values under the prior), and the outputs of MetaSim are each node’s new attributes (again normalized between 0 and 1). We use the GNN architecture of Kar *et al.* [9]: a three-layer GNN encoder with  $32 \rightarrow 64 \rightarrow 128$  features channels and a three-layer GNN decoder with  $128 \rightarrow 64 \rightarrow 32$  feature channels. Additionally, we use linear layers to encode and decode the per-node attributes.

Note that we train MetaSim using a supervised algorithm with heuristically generated ground truth scene graphs. In particular, given a real traffic scene, we first associate each actor to a lane segment; if this is not possible, the actor is not included in the scene graph. Next, we modify the attributes of each actor according to Prob. Grammar’s prior. Finally, this modified scene graph is given as input to MetaSim, and we train MetaSim to transform the modified attributes back to their original ones. In our setting, this training process was both faster and more stable than the original unsupervised algorithm.

**Procedural:** Our Procedural baseline is inspired by methods that operate directly on the road topology of the traffic scene [21, 20, 7, 13]. Specifically, given a *lane graph* of the scene [10], Procedural uses a set of rules to place actors onto lane centerlines. First, we determine a set of valid routes traversing the entire lane graph. Each valid route is a sequence of successive lane centerlines along which actors can traverse without violating traffic rules; *e.g.*, running red lights, merging onto an oncoming lane, *etc.* Next, we place actors onto each route such that successive actors maintain a random clearance (drawn from an exponential distribution) and no two actors collide. Each actor’s bounding box size is sampled from a Gaussian KDE fitted to the training dataset, and its heading is determined by the tangent vector along its lane centerline at its location. Finally, we initialize the speed of each actor such that successive actors maintain a random time gap (drawn from an exponential distribution). Procedural is similar to the heuristics underlying [21, 20, 7] but generalized to handle arbitrary road topologies. Similar to Prob. Grammar, Procedural can generate only vehicles and bicyclists since existing HD maps do not provide sidewalks. We believe this limitation highlights the difficulty of using a heuristics-based approach.

**Lane Graph:** Inspired by MetaSim, we also consider a learning-based version of Procedural. Specifically, given a traffic scene generated by Procedural, we use a lane graph neural network to transform the attributes of each actor; *i.e.*, location, bounding box size, heading, and velocity. Our lane graph neural network follows the design of the state-of-the-art motion forecasting model by Liang *et al.* [10]. It consists of MapNet for extracting map topology features and four fusion modules: actor-to-lane, lane-to-lane, lane-to-actor, and actor-to-actor. We train Lane Graph using heuristically generated ground truth, as in our MetaSim baseline.

**LayoutVAE:** Our implementation of LayoutVAE largely follows that of Jyothi *et al.* [8]. To adapt LayoutVAE to traffic scene generation, we first augment the original model with an additional CNN to extract map features. In particular, given a bird’s eye view multi-channel image of the HD map, we use the backbone architecture of Liang *et al.* [11] to extract multi-scale map features, which we subsequently average-pool into a feature vector. This is then given to LayoutVAE as input in place of the label set encoding used in the original setting<sup>2</sup>. Our second modification enables LayoutVAE to output oriented

<sup>2</sup>The label set in our setting is fixed to be vehicles, pedestrians, and bicyclists.

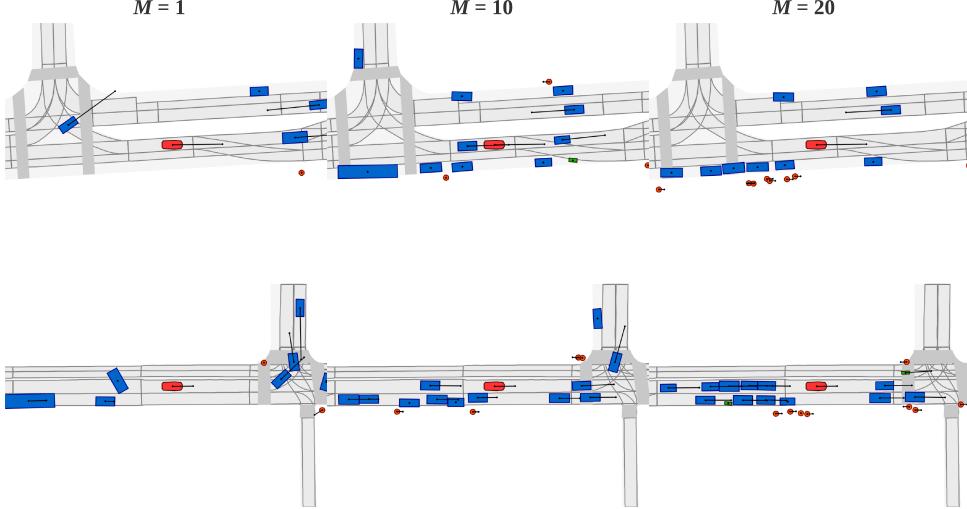


Figure 2: Traffic scenes generated by SceneGen using  $M = 1, 10, 20$  sample proposals for ATG4D.

bounding boxes and velocities. Specifically, we replace the spherical quadrivariate Gaussian distribution of its BBoxVAE with a bivariate Gaussian distribution for location, a bivariate log-normal distribution for bounding box size, and a bivariate Gaussian distribution for velocity. To evaluate the log-likelihood of a scene, we use Monte-Carlo approximation with 1000 samples from the conditional prior [8].

## 2.2. MMD Metrics

To complement our likelihood-based metric, we compute a sample-based metric as well: maximum mean discrepancy (MMD) [4]. As we discussed in the main text, MMD measures a distance between two distributions  $p$  and  $q$  as

$$\text{MMD}^2(p, q) = \mathbb{E}_{x, x' \sim p}[k(x, x')] + \mathbb{E}_{y, y' \sim q}[k(y, y')] - 2\mathbb{E}_{x \sim p, y \sim q}[k(x, y)] \quad (18)$$

for some kernel  $k$ . Following [24, 12], we compute MMD using Gaussian kernels (with bandwidth  $\sigma = 1$ ) with the total variation distance to compare scene statistics between generated and real traffic scenes. In particular, we first sample a set  $P$  of real traffic scenes from the evaluation dataset. Conditioned the SDV state and HD map of the scenes in  $P$ , we generate a set  $Q$  of synthetic scenes using the method under evaluation. Then, we approximate MMD as:

$$\text{MMD}^2(p, q) \approx \frac{1}{|P|^2} \sum_{x \in P} \sum_{x' \in P} k(x, x') + \frac{1}{|Q|^2} \sum_{y \in Q} \sum_{y' \in Q} k(y, y') - \frac{2}{|P||Q|} \sum_{x \in P} \sum_{y \in Q} k(x, y) \quad (19)$$

Our scene statistics measure the distribution of classes, bounding box sizes (in  $m^2$ ), speeds (in  $m/s$ ), and heading angles (relative to that of the SDV) for each scene. Empty scenes are discarded since these scene statistics are undefined. Since MMD is expensive to compute, in ATG4D, we form  $P$  by sampling the evaluation dataset by every 25th scene, yielding approximately 5000 scenes. We compute MMD over the full Argoverse validation set as it contains 5015 scenes only.

We also compute MMD in the feature space of a pre-trained motion forecasting model. This is similar to some popular metrics for evaluating generative models such IS [17], FID [5], and KID [2], except we use a motion forecasting model as our feature extractor. Here, our motion forecasting model takes a bird’s eye view multi-channel image of the actors in the scene and regresses the future locations of each actor over the next 3 seconds in 0.5s increments. We use the actor rasterization procedure described in Sec. 1.1 and the model architecture from [22], and we train the model using 4000 training log from the ATG4D training set. To obtain a feature vector summarizing the scene, we average pool the model’s backbone features along its spatial dimensions. Then, to compute MMD, we use the RBF kernel with bandwidth  $\sigma = 1$ .

$M$	Features	Class	Size	Speed	Heading
1	0.13	0.05	0.05	0.10	0.10
10	0.11	0.20	0.06	0.33	0.08
20	0.11	0.30	0.07	0.41	0.08

Table 2: Analysis of the number of sample proposals  $M$  on ATG4D. The reported numbers are the MMD metrics computed between distributions of features extracted by a motion forecasting model and various scene statistics (see main text).

### 3. Additional Experiment Results

#### 3.1. Vehicle MMD Metrics

In Tab. 1, we report vehicle-only MMD metrics for ATG4D and Argoverse. Specifically, we compute scene statistics for generated and real traffic scenes using vehicle actors only. As before, scenes with no vehicles are discarded during evaluation. This allows for an alternative comparison that controls for the class most easily handled by heuristics; *i.e.*, vehicles. Overall, we see that SceneGen still achieves the best results among the competing methods. This result reaffirms our claim that heuristics-based methods are insufficient to model the full complexity and diversity of real world traffic scenes.

#### 3.2. Sampling Strategy Analysis

As discussed in the main text, SceneGen uses a sampling strategy inspired by *nucleus sampling* [6]. Specifically, at each generation step, we sample each of SceneGen’s position, heading, and velocity distributions  $M$  times and return the most likely sample as output. In Tab. 2 and Fig. 2, we analyze the effects of using different numbers of sample proposals  $M = 1, 10, 20$ . We see that using  $M > 1$  decreases MMD on deep features, indicating that scene-level realism is improved. This improvement is even more evident in Fig. 2, where we see vehicles disregarding the rules of traffic when  $M = 1$ . With more fine-grained tuning of  $M$ , we expect to see improvements in the actor-level statistics as well; *i.e.*, class, size, and speed.

### 4. Additional Qualitative Results

In Fig. 3 and 4, we present an array of additional qualitative results for ATG4D and Argoverse respectively. Here, we compare traffic scenes generated by SceneGen, MetaSim, Lane Graph, and LayoutVAE. From these visualizations, we see that SceneGen generates traffic scenes that best reflect the complexity and diversity of real world traffic scenes. For example, in the second-to-last row of Fig. 3, we show a traffic scene generated by SceneGen in which a vehicle performs a three-point turn. In the bottom row of Fig. 3, we also show a scene in which two bicyclists perform an left turn using the car lane. These scenes highlight SceneGen’s ability to model rare but plausible traffic scenes that could occur in the real world.

In Fig. 5 and 6, we also showcase the diversity of traffic scenes that SceneGen is able to generate. Each row in the figures show four samples from our model when given the same SDV state and HD map as inputs. From these visualizations, we see that SceneGen captures the multi-modality of real world traffic scenes well. For example, the top row of Fig. 5 shows four traffic scenes generated for a four-way intersection. Here, we see samples in which pedestrians cross the intersection, vehicles perform an unprotected left turn, and a large bus goes straight.

Finally, in Fig. 7, we visualize the quantized location heatmaps for steps  $t = 0, 5, 10, 15, 20$  of the generation process. Each row shows the categorical distribution from which we sample the next actor’s location. From these visualizations, we see that SceneGen is able to model the distribution over actor locations (and the corresponding uncertainties) quite precisely. For example, the distribution over vehicle locations are concentrated around lane centerlines and the distribution over pedestrian locations are diffused over crosswalks and sidewalks.

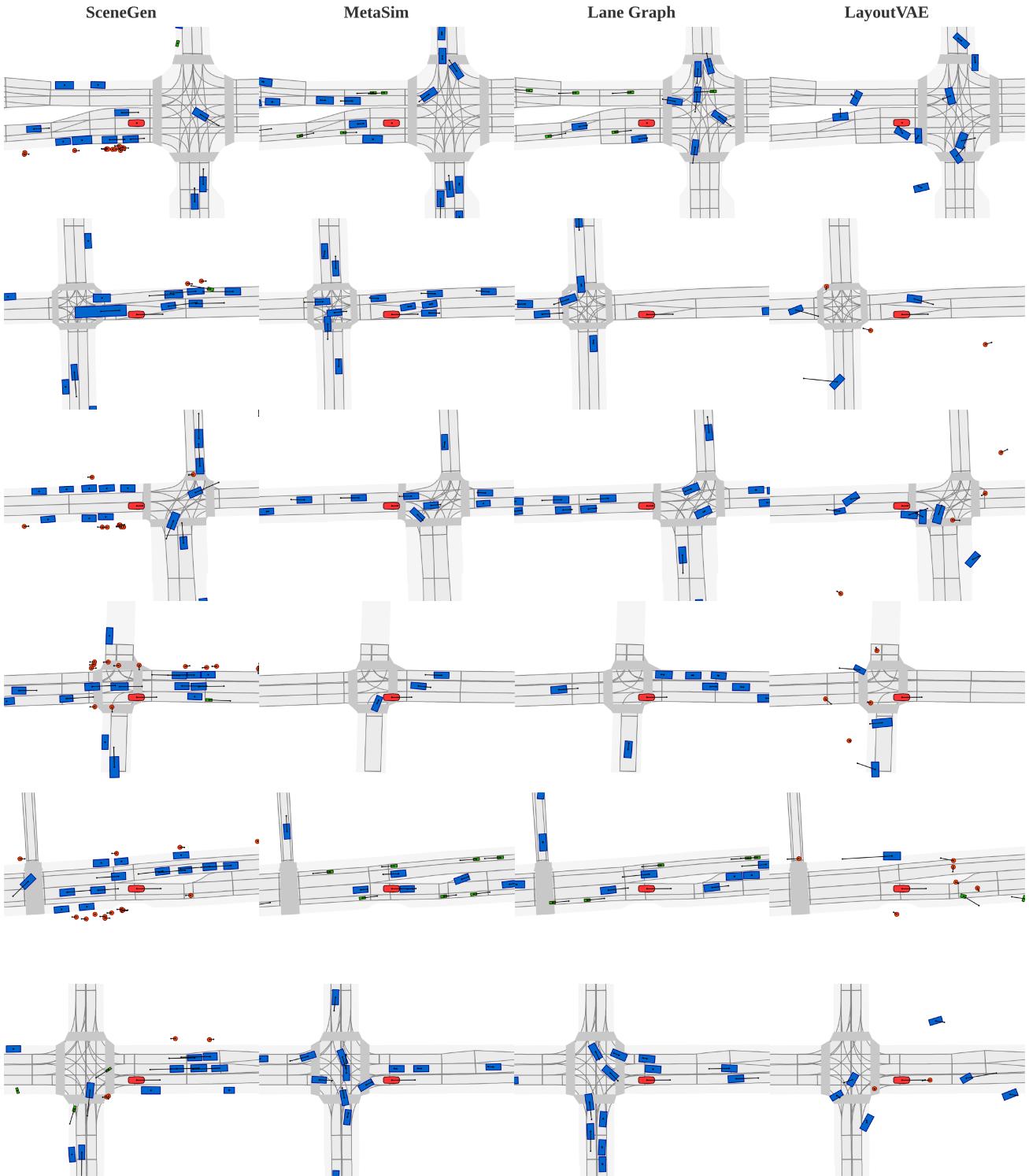


Figure 3: Qualitative comparison of traffic scenes generated by SceneGen and various baselines on ATG4D. The ego SDV is shown in red; vehicles in blue; pedestrians in orange; and bicyclists in green. We visualize lane segments and drivable surfaces in light grey and crosswalks in dark grey.

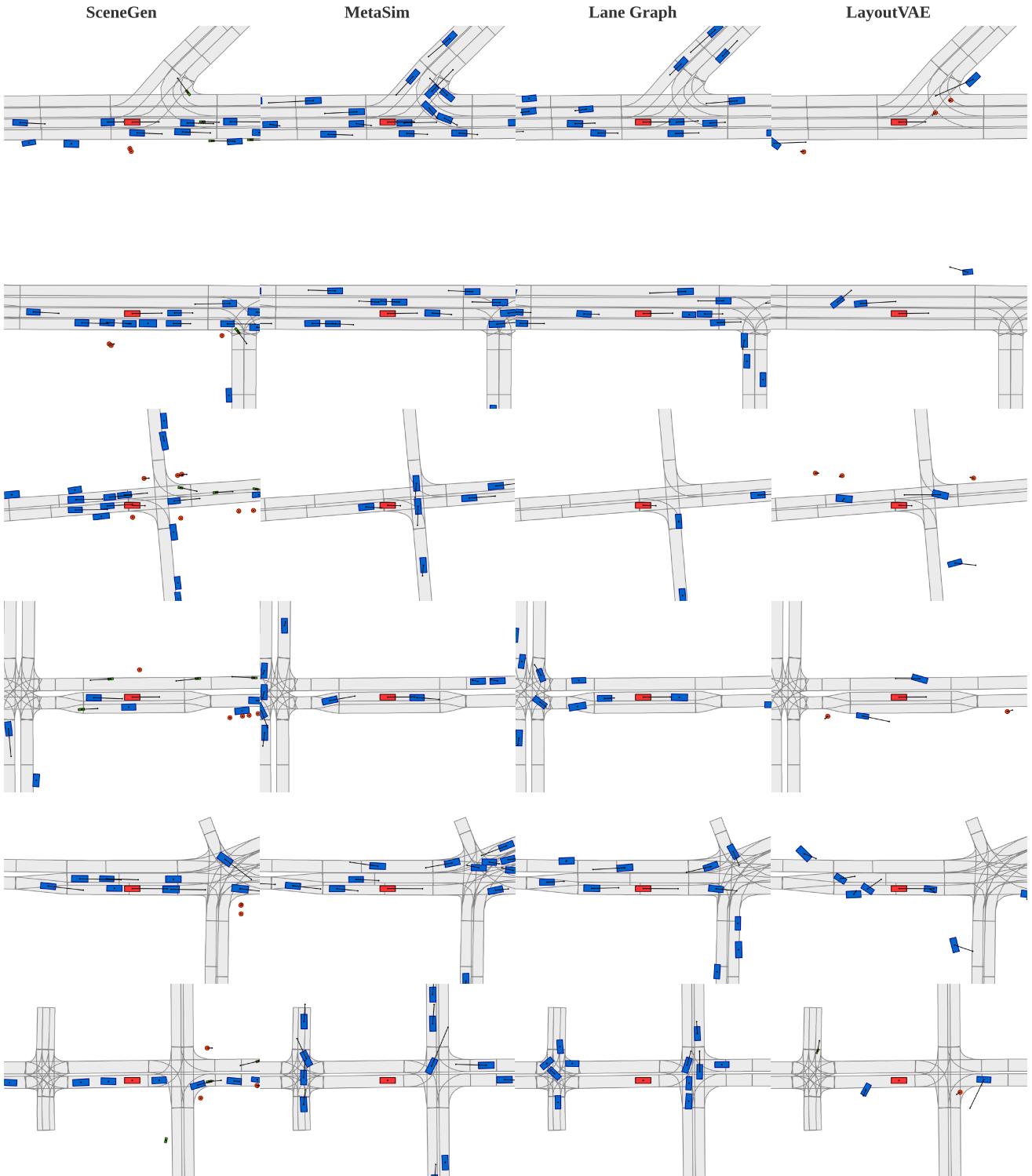


Figure 4: Qualitative comparison of traffic scenes generated by SceneGen and various baselines on Argoverse. The ego SDV is shown in red; vehicles in blue; pedestrians in orange; and bicyclists in green. We visualize lane segments in light grey.

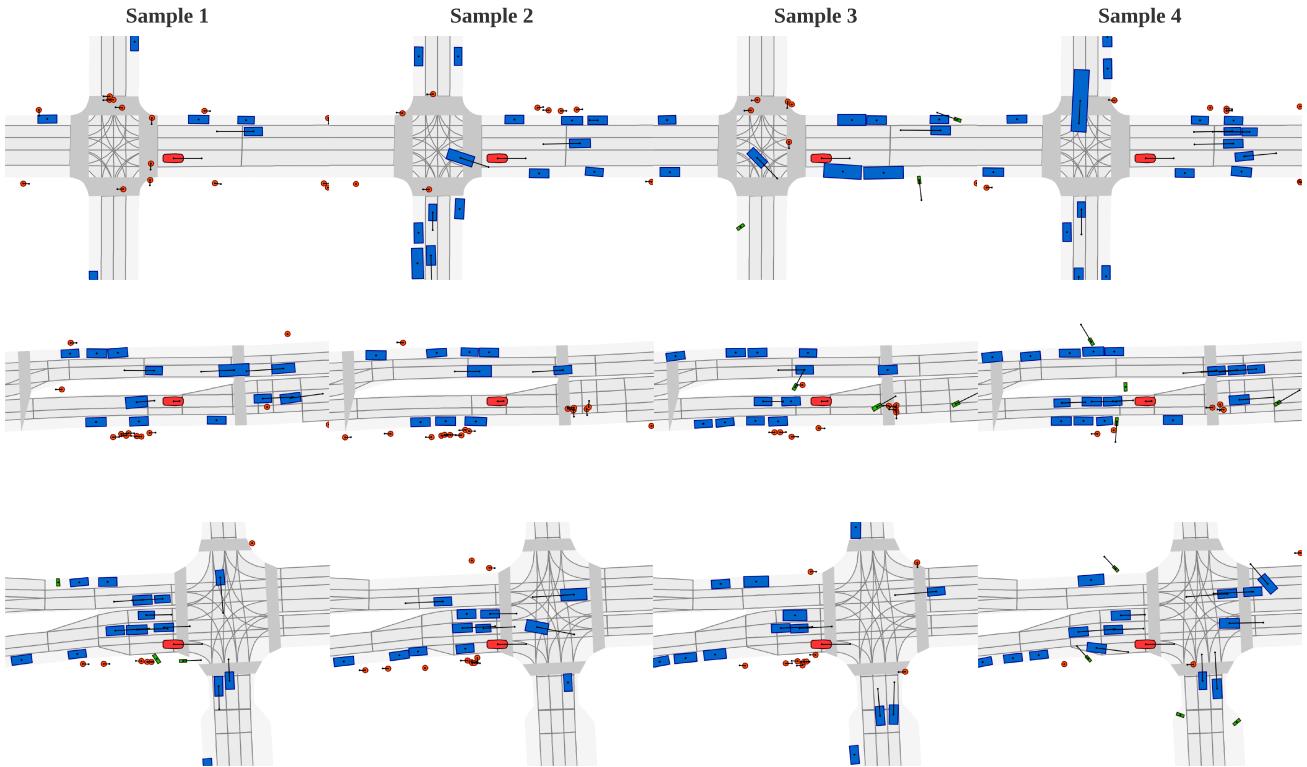


Figure 5: Traffic scenes generated by SceneGen on ATG4D. The traffic scenes in each row are generated from the same SDV state and HD map inputs. Each traffic scene is a distinct sample drawn from our model.

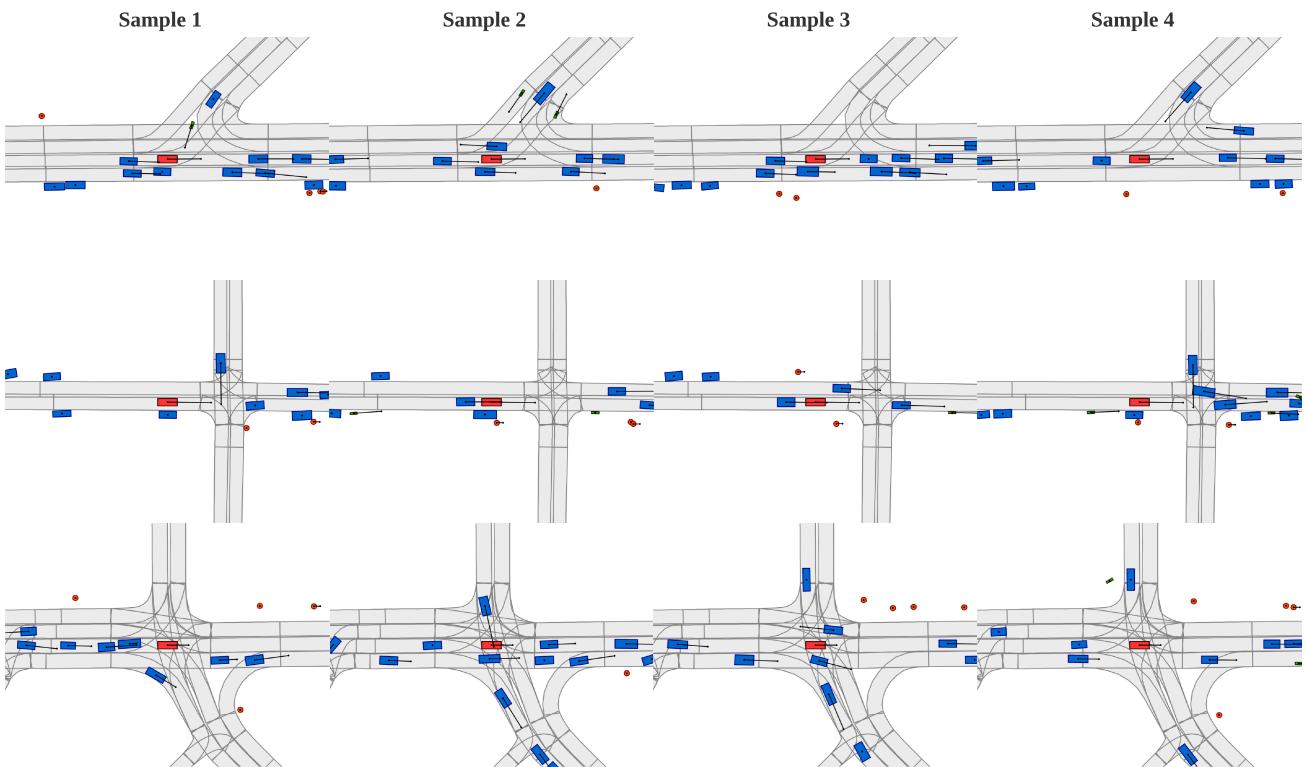


Figure 6: Traffic scenes generated by SceneGen on Argoverse. The traffic scenes in each row are generated from the same SDV state and HD map inputs. Each traffic scene is a distinct sample drawn from our model.

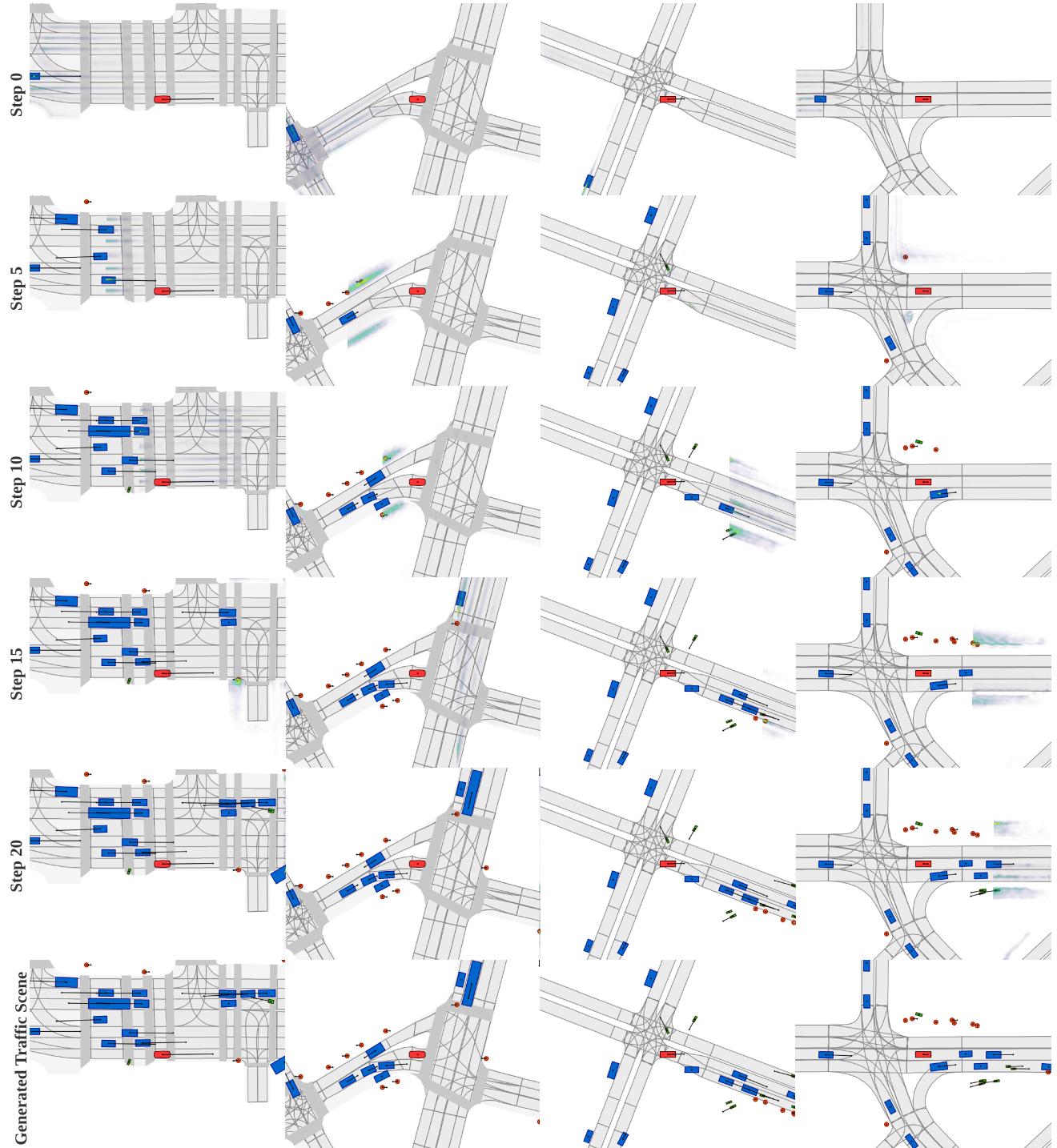


Figure 7: Traffic scenes generated by SceneGen on ATG4D (first two columns) and Argoverse (last two columns). We visualize the quantized location heatmap for steps  $t = 0, 5, 10, 15, 20$  of the generation process. Each column represents the generation process for one traffic scene. Bright yellow means higher likelihood.

## References

- [1] Mayank Bansal, Alex Krizhevsky, and Abhijit S. Ogale. Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst. In *RSS*, 2019. 1
- [2] Mikolaj Binkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD gans. In *ICLR*, 2018. 6
- [3] Jeevan Devarajan, Amlan Kar, and Sanja Fidler. Meta-sim2: Unsupervised learning of scene structure for synthetic data generation. 2020. 4
- [4] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander J. Smola. A kernel two-sample test. *JMLR*, 2012. 6
- [5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017. 6
- [6] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *ICLR*, 2020. 7
- [7] Stefan Jesenski, Jan Erik Stellet, Florian A. Schiegg, and J. Marius Zöllner. Generation of scenes in intersections for the validation of highly automated driving functions. In *IV*, 2019. 5
- [8] Akash Abdu Jyothi, Thibaut Durand, Jiawei He, Leonid Sigal, and Greg Mori. Layoutvae: Stochastic scene layout generation from a label set. In *ICCV*, 2019. 5, 6
- [9] Amlan Kar, Aayush Prakash, Ming-Yu Liu, Eric Cameracci, Justin Yuan, Matt Rusiniak, David Acuna, Antonio Torralba, and Sanja Fidler. Meta-sim: Learning to generate synthetic datasets. In *ICCV*, 2019. 4, 5
- [10] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *ECCV*, 2020. 5
- [11] Ming Liang, Bin Yang, Wenyuan Zeng, Yun Chen, Rui Hu, Sergio Casas, and Raquel Urtasun. Pnpnet: End-to-end perception and prediction with tracking in the loop. In *CVPR*, 2020. 5
- [12] Renjie Liao, Yujia Li, Yang Song, Shenlong Wang, William L. Hamilton, David Duvenaud, Raquel Urtasun, and Richard S. Zemel. Efficient graph generation with graph recurrent attention networks. In *NeurIPS*, 2019. 6
- [13] Sivabalan Manivasagam, Shenlong Wang, Kelvin Wong, Wenyuan Zeng, Mikita Sazanovich, Shuhan Tan, Bin Yang, Wei-Chiu Ma, and Raquel Urtasun. Lidarsim: Realistic lidar simulation by leveraging the real world. In *CVPR*, 2020. 5
- [14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 3
- [15] Aayush Prakash, Shaad Boochoon, Mark Brophy, David Acuna, Eric Cameracci, Gavriel State, Omer Shapira, and Stan Birchfield. Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In *ICRA*, 2019. 4
- [16] Sergey Prokudin, Peter V. Gehler, and Sebastian Nowozin. Deep directional statistics: Pose estimation with uncertainty quantification. In *ECCV*, 2018. 1, 3
- [17] Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, 2016. 6
- [18] Alexander Sergeev and Mike Del Balso. Horovod: fast and easy distributed deep learning in tensorflow. *arXiv*, 2018. 3
- [19] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NeurIPS*, 2015. 1
- [20] Tim Allan Wheeler and Mykel J. Kochenderfer. Factor graph scene distributions for automotive safety analysis. In *ITSC*, 2016. 5
- [21] Tim Allan Wheeler, Mykel J. Kochenderfer, and Philipp Robbel. Initial scene configurations for highway traffic propagation. In *ITSC*, 2015. 5
- [22] Kelvin Wong, Qiang Zhang, Ming Liang, Bin Yang, Renjie Liao, Abbas Sadat, and Raquel Urtasun. Testing the safety of self-driving vehicles by simulating perception and prediction. *ECCV*, 2020. 6
- [23] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, 2018. 2
- [24] Jiaxuan You, Rex Ying, Xiang Ren, William L. Hamilton, and Jure Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models. In *ICML*, 2018. 6