Supplementary for "StretchBEV: Stretching Future Instance Prediction Spatially and Temporally"

Adil Kaan Akan and Fatma Güney

KUIS AI Center, Koc University, Istanbul, Turkey
{kakan20, fguney}@ku.edu.tr
https://kuis-ai.github.io/stretchbev

Abstract. In this part, we provide additional illustrations, derivations, and results for our paper "StretchBEV: Stretching Future Instance Prediction Spatially and Temporally". We first show the full derivation of the Evidence Lower Bound (ELBO) in Section 1. In Section 2, we explain the architectural choices and training details. We present the detailed versions of the quantitative results in the main paper. In addition, we present more ablation experiments with the content variable and perform a comparison in terms of inference speed. In Section 4, we provide more qualitative results comparing our method to ground truth and FIERY [4], and also visualizations of samples for diversity. Video visualizations are available at our website.

1 Evidence Lower Bound

In this section, we derive the variational lower bound for the proposed model following [2]. The changes in our derivation are mainly due to excluding the content variable and including the output modalities in the derivations.

Using the original variational lower bound of variational autoencoders [6] in (1):

$$\begin{aligned}
&\log p(\mathbf{s}_{1:T}, \mathbf{o}_{1:T}) \\
&\quad \geq \mathbb{E}_{(\tilde{\mathbf{z}}_{2:T}, \tilde{\mathbf{y}}_{1:T}) \sim q_{Z,Y}} \log p(\mathbf{s}_{1:T}, \mathbf{o}_{1:T} | \tilde{\mathbf{z}}_{2:T}, \tilde{\mathbf{y}}_{1:T}) - D_{\mathrm{KL}}(q_{Z,Y} \mid\mid p(\mathbf{y}_{1:T}, \mathbf{z}_{2:T})) \qquad (1) \\
&\quad = \mathbb{E}_{(\tilde{\mathbf{z}}_{2:T}, \tilde{\mathbf{y}}_{1:T}) \sim q_{Z,Y}} \log p\left(\mathbf{s}_{1:T}, \mathbf{o}_{1:T} | \tilde{\mathbf{z}}_{2:T}, \tilde{\mathbf{y}}_{1:T}\right) \qquad (2) \\
&\qquad\qquad - D_{\mathrm{KL}}(q\left(\mathbf{y}_1, \mathbf{z}_{2:T} | \mathbf{s}_{1:T}, \mathbf{o}_{1:T}\right) \mid\mid p(\mathbf{y}_{1:T}, \mathbf{z}_{2:T})) \\
&\quad = \mathbb{E}_{(\tilde{\mathbf{z}}_{2:T}, \tilde{\mathbf{y}}_{1:T}) \sim q_{Z,Y}} \sum_{t=1}^{T} \log p(\mathbf{s}_t | \tilde{\mathbf{y}}_t) + \log p(\mathbf{o}_t | \mathbf{s}_t) \qquad (3) \\
&\qquad\qquad - D_{\mathrm{KL}}(q(\mathbf{y}_1, \mathbf{z}_{2:T} | \mathbf{s}_{1:T}, \mathbf{o}_{1:T}) \mid\mid p(\mathbf{y}_{1:T}, \mathbf{z}_{2:T}))
\end{aligned}$$

where:

- (2) is given by the forward and inference models factorizing p and q in Equations (3,4,5) in the main paper.

- The $\mathbf{y}_{2:T}$ variables are deterministic functions of $\mathbf{y}_1$ and $\mathbf{z}_{2:T}$ with respect to p and q;
- (3) results from the factorization of $p(\mathbf{s}_{1:T}|\mathbf{y}_{1:T}, \mathbf{z}_{1:T})$ in Equation (3) in the main paper.
- $\log\, p(\mathbf{o}_t|\mathbf{s}_t)$ is also deterministic and corresponds to supervised decoding of output modalities (Sec 3.4 in the main paper).

From there, by using the integral formulation of D_{KL}:

$$\begin{aligned}
&\log\, p(\mathbf{s}_{1:T}, \mathbf{o}_{1:T}) \\
&\geq \mathbb{E}_{(\tilde{\mathbf{z}}_{2:T}, \tilde{\mathbf{y}}_{1:T}) \sim q_{Z,Y}} \sum_{t=1}^{T} \log\, p(\mathbf{s}_t, \mathbf{o}_t|\tilde{\mathbf{y}}_t) \\
&+ \int \cdots \int_{\mathbf{y}_1, \mathbf{z}_{2:T}} q(\mathbf{y}_1, \mathbf{z}_{2:T}|\mathbf{s}_{1:T}, \mathbf{o}_{1:T}) \log \frac{p(\mathbf{y}_1, \mathbf{z}_{2:T})}{q(\mathbf{y}_1, \mathbf{z}_{2:T}|\mathbf{s}_{1:T}, \mathbf{o}_{1:T})} \mathrm{d}\mathbf{z}_{2:T} \mathrm{d}\mathbf{y}_1
\end{aligned} \tag{4}$$

$$\begin{aligned}
&= \mathbb{E}_{(\tilde{\mathbf{z}}_{2:T}, \tilde{\mathbf{y}}_{1:T}) \sim q_{Z,Y}} \sum_{t=1}^{T} \log\, p(\mathbf{s}_t|\tilde{\mathbf{y}}_t) + \log\, p(\mathbf{o}_t|\mathbf{s}_t) - D_{\mathrm{KL}}(q(\mathbf{y}_1|\mathbf{s}_{1:T}) \;||\; p(\mathbf{y}_1)) \\
&+ \mathbb{E}_{\tilde{\mathbf{y}}_1 \sim q(\mathbf{y}_1|\mathbf{s}_{1:T})} \left[\int \cdots \int_{\mathbf{z}_{2:T}} q(\mathbf{z}_{2:T}|\mathbf{s}_{1:T}, \mathbf{o}_{1:T}, \tilde{\mathbf{y}}_1) \log \frac{p(\mathbf{z}_{2:T}|\tilde{\mathbf{y}}_1)}{q(\mathbf{z}_{2:T}|\mathbf{s}_{1:T}, \mathbf{o}_{1:T}, \tilde{\mathbf{y}}_1)} \mathrm{d}\mathbf{z}_{2:T} \right]
\end{aligned} \tag{5}$$

$$\begin{aligned}
&= \mathbb{E}_{(\tilde{\mathbf{z}}_{2:T}, \tilde{\mathbf{y}}_{1:T}) \sim q_{Z,Y}} \sum_{t=1}^{T} \log\, p(\mathbf{s}_t|\tilde{\mathbf{y}}_t) + \log\, p(\mathbf{o}_t|\mathbf{s}_t) - D_{\mathrm{KL}}(q(\mathbf{y}_1|\mathbf{s}_{1:k}) \;||\; p(\mathbf{y}_1)) \\
&+ \mathbb{E}_{\tilde{\mathbf{y}}_1 \sim q(\mathbf{y}_1|\mathbf{s}_{1:k})} \left[\int \cdots \int_{\mathbf{z}_{2:T}} q(\mathbf{z}_{2:T}|\mathbf{s}_{1:T}, \mathbf{o}_{1:T}, \tilde{\mathbf{y}_1}) \log \frac{p(\mathbf{z}_{2:T}|\tilde{\mathbf{y}}_1)}{q(\mathbf{z}_{2:T}|\mathbf{s}_{1:T}, \mathbf{o}_{1:T}, \tilde{\mathbf{y}}_1)} \mathrm{d}\mathbf{z}_{2:T} \right]
\end{aligned} \tag{6}$$

$$\begin{aligned}
&= \mathbb{E}_{(\tilde{\mathbf{z}}_{2:T}, \tilde{\mathbf{y}}_{1:T}) \sim q_{Z,Y}} \sum_{t=1}^{T} \log\, p(\mathbf{s}_t|\tilde{\mathbf{y}}_t) + \log\, p(\mathbf{o}_t|\mathbf{s}_t) - D_{\mathrm{KL}}(q(\mathbf{y}_1|\mathbf{s}_{1:k}) \;||\; p(\mathbf{y}_1)) \\
&+ \mathbb{E}_{\tilde{\mathbf{y}}_1 \sim q(\mathbf{y}_1|\mathbf{s}_{1:k})} \left[\int \cdots \int_{\mathbf{z}_{2:T}} \prod_{t=2}^{T} q(\mathbf{z}_t|\mathbf{s}_{1:t}, \mathbf{o}_{1:t}) \sum_{t=2}^{T} \log \frac{p(\mathbf{z}_t|\tilde{\mathbf{y}}_1, \mathbf{z}_{2:t-1})}{q(\mathbf{z}_t|\mathbf{s}_{1:t}, \mathbf{o}_{1:t})} \mathrm{d}\mathbf{z}_{2:T} \right]
\end{aligned} \tag{7}$$

$$\begin{aligned}
&= \mathbb{E}_{(\tilde{\mathbf{z}}_{2:T}, \tilde{\mathbf{y}}_{1:T}) \sim q_{Z,Y}} \sum_{t=1}^{T} \log\, p(\mathbf{s}_t|\tilde{\mathbf{y}}_t) + \log\, p(\mathbf{o}_t|\mathbf{s}_t) - D_{\mathrm{KL}}(q(\mathbf{y}_1|\mathbf{s}_{1:k}) \;||\; p(\mathbf{y}_1)) \\
&- \mathbb{E}_{\tilde{\mathbf{y}}_1 \sim q(\mathbf{y}_1|\mathbf{s}_{1:k})} D_{\mathrm{KL}}(q(\mathbf{z}_2|\mathbf{s}_{1:t}, \mathbf{o}_{1:t}) \;||\; p(\mathbf{z}_2|\tilde{\mathbf{y}}_1)) \\
&+ \mathbb{E}_{\tilde{\mathbf{y}}_1 \sim q(\mathbf{y}_1|\mathbf{s}_{1:k})} \mathbb{E}_{\tilde{\mathbf{z}}_2 \sim q(\mathbf{z}_2|\mathbf{s}_{1:2}, \mathbf{o}_{1:2})} \\
&\qquad \left[\int \cdots \int_{\mathbf{z}_{3:T}} \prod_{t=3}^{T} q(\mathbf{z}_t|\mathbf{s}_{1:t}, \mathbf{o}_{1:t}) \sum_{t=3}^{T} \log \frac{p(\mathbf{z}_t|\tilde{\mathbf{y}}_1, \mathbf{z}_{2:t-1})}{q(\mathbf{z}_t|\mathbf{s}_{1:t}, \mathbf{o}_{1:t})} \mathrm{d}\mathbf{z}_{3:T} \right]
\end{aligned} \tag{8}$$

where:

- (6) follows from the inference model of Equation (5) in the main paper, where $\mathbf{y}_1$ only depends on $\mathbf{s}_{1:k}$;
- (7) is obtained from the factorizations of Equations (3,4,5) in the main paper.

By iterating (8)'s step on $\mathbf{z}_3, \ldots, \mathbf{z}_T$ and factorizing all expectations, we obtain:

$$\begin{aligned}
\log \quad & p(\mathbf{s}_{1:T}, \mathbf{o}_{1:T}) \qquad (9) \\
\geq \quad & \mathbb{E}_{(\tilde{\mathbf{z}}_{2:T}, \tilde{\mathbf{y}}_{1:T}) \sim q_{Z,Y}} \sum_{t=1}^{T} \log\, p(\mathbf{s}_t|\tilde{\mathbf{y}}_t) + \log\, p(\mathbf{o}_t|\mathbf{s}_t) - D_{\mathrm{KL}}(q(\mathbf{y}_1|\mathbf{s}_{1:k}) \;||\; p(\mathbf{y}_1)) \\
- \quad & \mathbb{E}_{\tilde{\mathbf{y}}_1 \sim q(\mathbf{y}_1|\mathbf{s}_{1:k})} \left(\mathbb{E}_{\tilde{\mathbf{z}}_t \sim q(\mathbf{z}_t|\mathbf{s}_{1:t}, \mathbf{o}_{1:t})} \right)_{t=2}^{T} \sum_{t=2}^{T} D_{\mathrm{KL}}(q(\mathbf{z}_t|\mathbf{s}_{1:t}, \mathbf{o}_{1:t}) \;||\; p(\mathbf{z}_t|\tilde{y}_1, \tilde{\mathbf{z}}_{1:t-1}))
\end{aligned}$$

and we finally retrieve Evidence Lower Bound in (6) in the main paper by using the factorization in (5) in the main paper:

$$\begin{aligned}
\log \quad & p(\mathbf{s}_{1:T}, \mathbf{o}_{1:T}) \qquad (10) \\
\geq \quad & \mathbb{E}_{(\tilde{\mathbf{z}}_{2:T}, \tilde{\mathbf{y}}_{1:T}) \sim q_{Z,Y}} \sum_{t=1}^{T} \log\, p(\mathbf{s}_t|\tilde{\mathbf{y}}_t) + \log\, p(\mathbf{o}_t|\mathbf{s}_t) - D_{\mathrm{KL}}(q(\mathbf{y}_1|\mathbf{s}_{1:k}) \;||\; p(\mathbf{y}_1)) \\
- \quad & \mathbb{E}_{(\tilde{\mathbf{z}}_{2:T}, \tilde{\mathbf{y}}_{1:T}) \sim q_{Z,Y}} \sum_{t=2}^{T} D_{\mathrm{KL}}(q(\mathbf{z}_t|\mathbf{s}_{1:t}, \mathbf{o}_{1:t}) \;||\; p(\mathbf{z}_t|\tilde{\mathbf{y}}_{t-1}))
\end{aligned}$$

2 Model and Training Details

In this section, we provide the details of the architectures used (Section 2.1), and the details of the training including the hyper-parameters used in the optimization (Section 2.2).

2.1 Model Details

Our models use the same framework as FIERY [4] following the same input-output setting to be comparable. Both models process $n = 6$ camera images at $(H_{\text{in}}, W_{\text{in}}) = (224 \times 480)$ for k conditioning time steps, i.e. $k = 3$, which results in 18 images in total. The minimum depth value we consider is $D_{\text{min}} = 2.0$m, which corresponds to the spatial extent of the ego-car. The maximum depth value is $D_{\text{max}} = 50.0$m, and the size of each depth slice is set to $D_{\text{size}} = 1.0$m. Our model uses the same bird's-eye view (BEV) encoder and future instance segmentation and motion decoder as FIERY [4]. For further details, we direct reviewers into their appendix section. Next, we explain the details of each block in our model and for the missing or unclear parts, the code is attached with the submission. We will also share the code and the trained models upon publication.

Dow-sampling Encoder and Up-sampling Decoder: Our model uses another encoder-decoder pair to reduce spatial size of feature extracted by the

BEV encoder. Down-sampling encoder contains 10 convolutional layers followed by batch normalization and Leaky ReLU activation. After 2 convolutional layers, we apply a dropout with probability of 0.25. At the end, we apply another convolutional layer with a batch normalization but with *tanh* activation at the end. Down-sampling encoder uses max-pooling after the second and fourth convolutions to reduce the spatial size to 1/4th resolution. Up-sampling decoder is the symmetric version of the down-sampling decoder. We use the "nearest" mode up-sampling instead of the max-pooling to increase the spatial size.

The First Latent State: We encode the conditioning frames with a small CNN to learn the first latent state $\mathbf{y}_1$. The network contains 4 convolutions followed by batch normalization and Leaky ReLU activation. We also add a Squeeze and Excitation layer after the second and the fourth convolution to enhance the learned features. At the end, we apply a convolutional layer which outputs $\boldsymbol{\mu}_\phi^{\mathbf{y}}$ and $\boldsymbol{\sigma}_\phi^{\mathbf{y}}$, and then we use them to sample the first state, $\mathbf{y}_1$.

Prior Distribution: We use another CNN to learn a prior distribution from the previous latent state $\mathbf{y}_{t-1}$. The network is the same as the first latent state network except for the input, we feed the previous latent variable, $\mathbf{y}_{t-1}$ at time t and it produces $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\sigma}_\theta$.

Posterior Distribution: For posterior distribution, we use recurrent neural network, GRU-Conv, which is a combination of SpatialGRUs and convolutions. Our goal is to learn a posterior distribution, $\boldsymbol{\mu}_\phi^{\mathbf{z}}$ and $\boldsymbol{\sigma}_\phi^{\mathbf{z}}$, representing the temporal dynamics. We first process image features extracted by the BEV encoder with our GRU-Conv network. Then, for each time step, we use the same network as the prior distribution to sample a posterior distribution. GRU-Conv contains 2 SpatialGRUs followed by 2 convolutional blocks, each of which contains 2 convolutions with 1×1 and 3×3 kernel sizes.

Dynamics Update: We use a network to update intermediate latent variables $\mathbf{y}_t$. We feed the previous latent variable $\mathbf{y}_{t-1}$ and the corresponding stochastic variable $\mathbf{z}_t$ at time t, and the output of the network is added to the previous latent variable, $\mathbf{y}_{t-1}$. The architecture is the same as the prior distribution architecture except that it only inputs one set of parameters at the end instead of two.

2.2 Training Details

We train our models with 2 V100 GPUs for 25 epochs at most. We will release all the scripts used for training and the checkpoints of the models used for evaluation.

Pre-training: Our models can be pre-trained to learn the dynamics update in an unsupervised manner. We simply initialize the BEV encoder from a checkpoint trained to segment the present time objects in bird's-eye view [7]. Moreover, we remove the future instance segmentation and the motion decoder. In this setting, our dynamics model learns to predict the BEV features of future time steps that are extracted by the BEV encoder conditioned on the features of the previous time steps. This way, our model learns to predict the future in the feature space without ground truth segmentation or motion. According to our results, the pre-training improves the results significantly for StretchBEV. We cannot do unsupervised pre-training for StretchBEV-P because it needs the ground truth labels in the posterior distribution.

Training Parameters: We train all our models for 25 epochs at most. We use a held-out validation set for model selection. We use the maximum batch size that fits into V100 GPUs, which is 2 for training and 12 for pre-training. We use 3×10^{-4} as a starting learning rate. We apply learning rate decay if the validation loss does not decrease for some threshold, which is the reason for a varying number of epochs depending on the model.

3 Detailed Quantitative Results

In this section, we provide extended versions of the tables in the main paper.

In Table 2, we provide our ablation table by also adding the results with the content variable. As explained in the main paper, content variable does not improve the performance of our models in contrast to the state of the art video prediction [2], therefore omitted in our formulation.

In Table 3, we provide the results of the future segmentation performances in the FISHING setting as proposed in [3]. Differently from the main paper, this table includes the result of StretchBEV as well. Both StretchBEV and StretchBEV-P outperform FIERY [4] in this setting.

In Table 4, we provide the quantitative results of Fig. 3 in the main paper, which shows the performance comparison over different temporal horizons. As can be seen from the table, our models StretchBEV and StretchBEV-P outperform FIERY [4] in far regions, especially StretchBEV-P by a large margin in terms of VPQ. The performance of StretchBEV is impressive in near IoU in longer settings.

In Table 5, we provide Generalized Energy Distance (GED) for both of our models StretchBEV and StretchBEV-P compared to FIERY [4]. As can be seen from the table, both our models are more diverse than FIERY in terms of GED. This shows the significance of modelling future uncertainty with time-independent stochastic latent variables.

Run-time and Parameter Comparison: We compare the inference speed of our model StretchBEV-P and FIERY [4] by measuring the average time needed to process a validation example in inference over 250 forward passes. Both models have almost the same inference speed (FIERY: 0.6436 seconds/example vs. StretchBEV-P: 0.6469 seconds/example). Although our model processes each time step separately, it does not introduce any drawbacks in terms of speed and its inference speed is almost the same as FIERY.

FIERY [5] has 8.1M parameters whereas StretchBEV-P has 16.2M. However, the runtime performances are still the same (0.64 sec per sample) thanks to the separation of generation from the learning of dynamics in our model.

4 Additional Visualizations and Qualitative Results

4.1 Qualitative Comparisons

In this subsection, we provide additional qualitative comparisons with FIERY [4] by also visualizing the ground truth. As it can be seen from the Figures 1, 2, 3, both models sometimes cannot capture the real trajectory, however, they have the notion of the objects and the motion in the scene. Most of the time, our model detects and segments the objects more correctly. Moreover, the trajectories generated by our model are more realistic and closer to the ground truth.

4.2 Measuring the Uncertainty Qualitatively

In Figures 4, 5, 6, we visualize three samples from FIERY and three samples from our model StretchBEV-P for the same scene in a row. When the scene is dynamic with moving vehicles, our model can generate more diverse samples by changing the speed of the vehicles. The samples generated by our model do not differ when the scene contains static objects which do not move. NuScenes dataset [1] contains a lot of vehicles that are parked and do not move throughout the scene. Therefore, our model does not learn to change the position of a static vehicle. In longer sequences, not only prediction but also tracking becomes harder, for example due to ID switches as can be seen in one of the samples (Fig. 6, row 5)

	Pre-training	Posterior w/labels	IoU (↑)		VPQ (↑)	
			Near	Far	Near	Far
StretchBEV	—	—	54.0 (53.3) ± 0.40	36.2 (35.8 ± 0.3)	43.6 (41.7 ± 1.1)	27.4 (26.0 ± 0.8)
	✓		56.2 (55.5 ± 0.4)	37.6 (37.1 ± 0.3)	47.9 (46.2 ± 1.0)	30.3 (29.0 ± 0.7)
StretchBEV-Det	✓	—	48.8 (48.8 ± 0)	32.1 (32.1 ± 0)	39.6 (39.6 ± 0)	23.6 (23.6 ± 0)
StretchBEV-MCD	✓	—	50.3 (48.2 ± 1.25)	32.8 (31.6 ± 0.7)	42.4 (39.0 ± 1.9)	25.3 (23.3 ± 1.2)
FIERY [4]	—	✓	**59.4**	36.7	50.2	29.4
Reproduced			59.0 (**58.8** ± 0.2)	35.9 (35.8 ± 0.1)	51.2 (50.5 ± 0.4)	29.5 (29.0 ± 0.3)
StretchBEV-P	—	✓	**60.6** (58.2 ± **1.4)**	**54.0 (52.5 ± 0.9)**	**56.6 (52.4 ± 2.4)**	**51.0 (48.3 ± 1.6)**

Table 1: **Ablation Study.** In this table, we present the results for the two versions of our model with (StretchBEV-P) and without (StretchBEV) using the labels for the output modalities in the posterior while learning the temporal dynamics and also show the effect of pre-training for the latter in comparison to FIERY [4] and our reproduced version of their results (Reproduced).

	Pre-training	Posterior w/labels	Content	IoU (↑)		VPQ (↑)	
				Near	Far	Near	Far
StretchBEV	—	—	—	53.3	35.8	41.7	26.0
	—		✓	51.9	34.1	40.8	25
	✓		—	55.5	37.1	46.0	29.0
FIERY [4]	—	✓	—	**59.4**	36.7	50.2	29.4
Reproduced				58.8	35.8	50.5	29.0
StretchBEV-P	—	✓	—	58.1	**52.5**	**53.0**	**47.5**
StretchBEV-P	—	✓	✓	57.6	51.9	51.5	46.8

Table 2: **Ablation Study.** Different than the table that we provide in the main paper, this table includes results with Content variable. As we stated before, content variable does not improve the results because most of the details are suppressed in the BEV representation.

Fishing-Cam [3]	Fishing-LiDAR [3]	FIERY [4]	StretchBEV	StretchBEV-P
30.0	44.3	57.3	58.8	**65.7**

Table 3: **Comparison of Semantic Segmentation Prediction.** In this table, we compare the predictions of our models, StretchBEV and StretchBEV-P for semantic segmentation to other BEV prediction methods in terms of IoU using the setting proposed in [3], i.e. 32.0m × 19.2m at 10cm resolution over 2s future.

	Short				Mid				Long			
	IoU (↑)		VPQ (↑)		IoU (↑)		VPQ (↑)		IoU (↑)		VPQ (↑)	
	Near	Far	Near	Far	Near	Far	Near	Far	Near	Far	Near	Far
StretchBEV	55.5	37.1	46.0	29.0	**47.7**	32.5	39.1	23.8	**43.7**	28.4	36.4	21.0
FIERY [4]	**58.8**	35.8	50.5	29.0	47.4	30.1	40.6	23.6	41.8	26.7	36.6	20.9
StretchBEV-P	58.1	**52.5**	**53.0**	**47.5**	46.8	**32.7**	**43.7**	**38.4**	38.2	**31.8**	**37.4**	**30.8**

Table 4: **Evaluation over Different Temporal Horizons.** This table extends Figure 3 in the main paper with the results of our models in comparison to FIERY [4] over short (2.0s), mid (4.0s), and long (6.0s) temporal horizons.

	Generalized Energy Distance (↓)					
	Short		Mid		Long	
	Near	Far	Near	Far	Near	Far
FIERY [4]	106.09	140.36	118.74	147.26	127.18	152.38
StretchBEV	103.97	132.38	114.11	138.15	119.01	142.51
StretchBEV-P	**82.04**	**85.51**	**94.02**	**98.45**	**101.90**	**109.12**

Table 5: **Quantitative Evaluation of Diversity.** This table compares the results of our models to the reproduced results of FIERY [4] in terms of Generalized Energy Distance based on VPQ (lower better) for evaluating diversity.

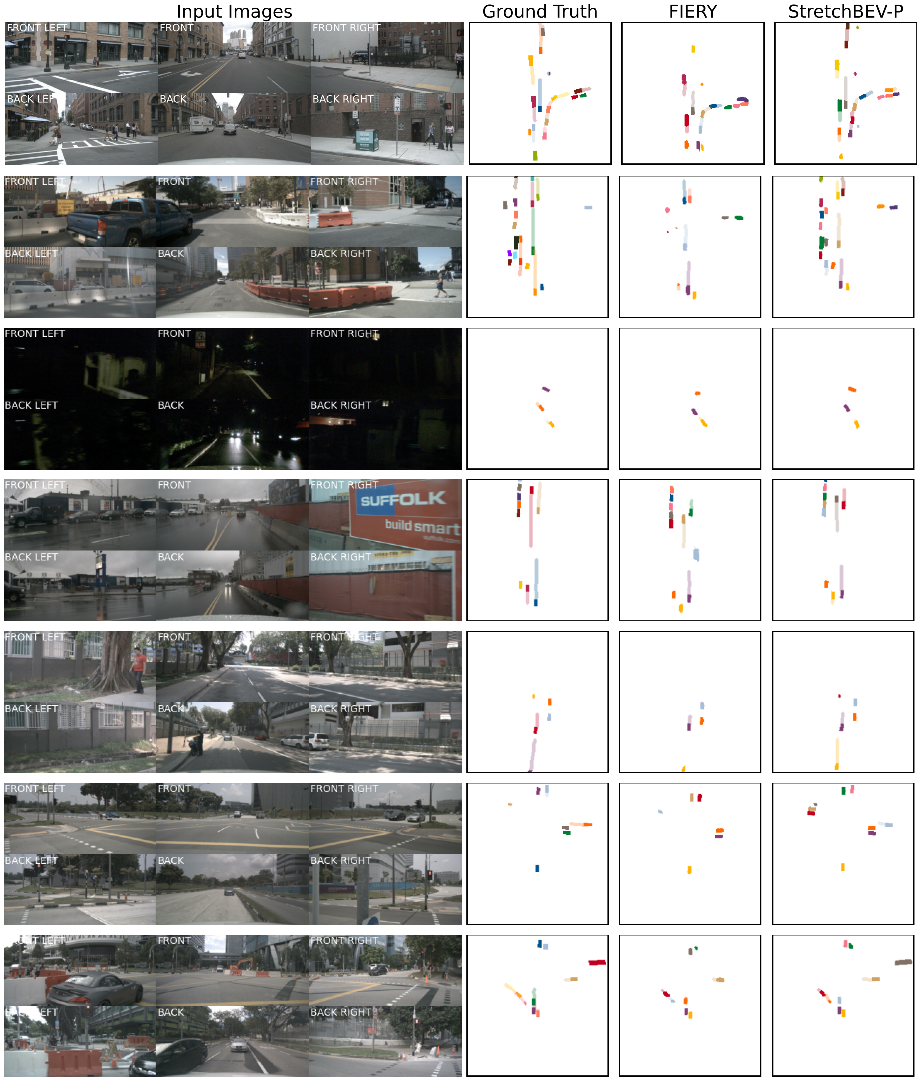

Fig. 1: **Qualitative Comparison for 2 seconds into future.** In this figure, we qualitatively compare the results of our model StretchBEV-P **(right)** to the ground truth **(left)** and FIERY [4] **(middle)** over short temporal horizon, which corresponds to predicting 2.0 seconds into the future. Each color represents an instance of a vehicle with its trajectory trailing in the same color transparently.

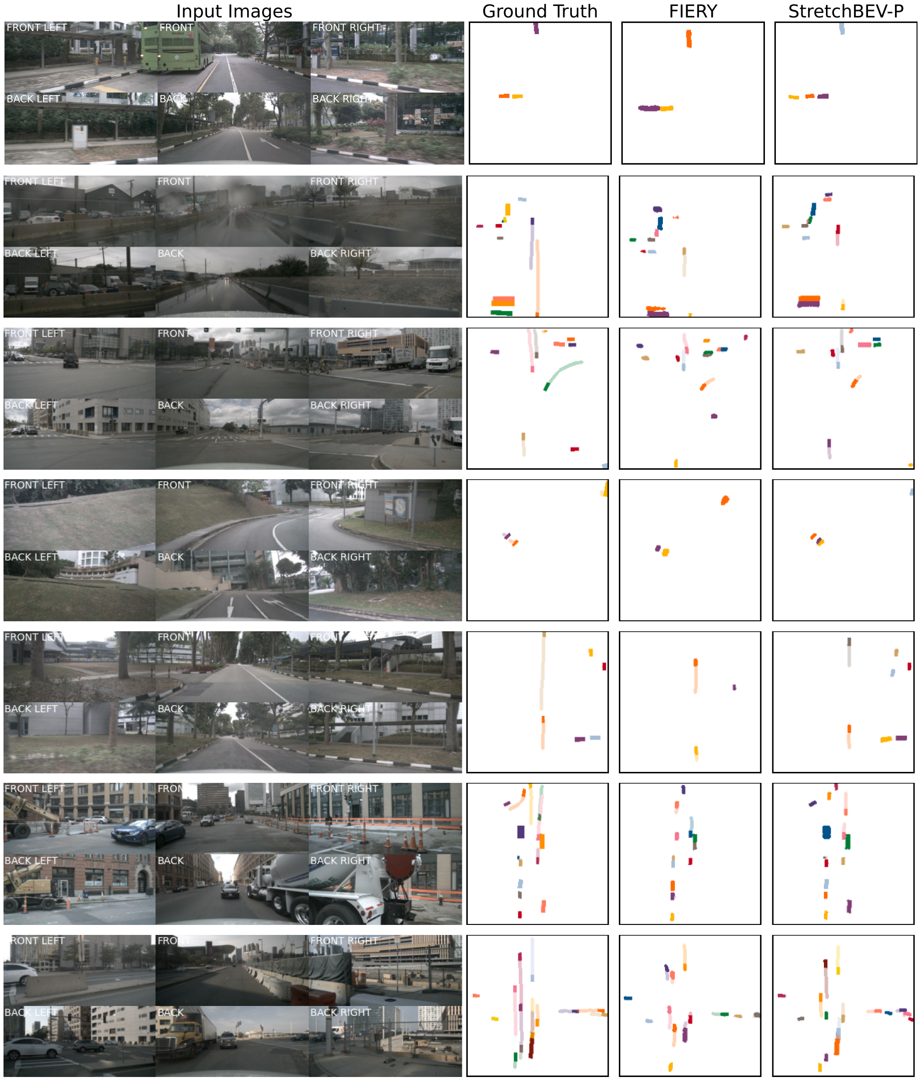

Fig. 2: **Qualitative Comparison for 4 seconds into future.** In this figure, we qualitatively compare the results of our model StretchBEV-P **(right)** to the ground truth **(left)** and FIERY [4] **(middle)** over mid temporal horizon, which corresponds to predicting 4.0 seconds into the future. Each color represents an instance of a vehicle with its trajectory trailing in the same color transparently.

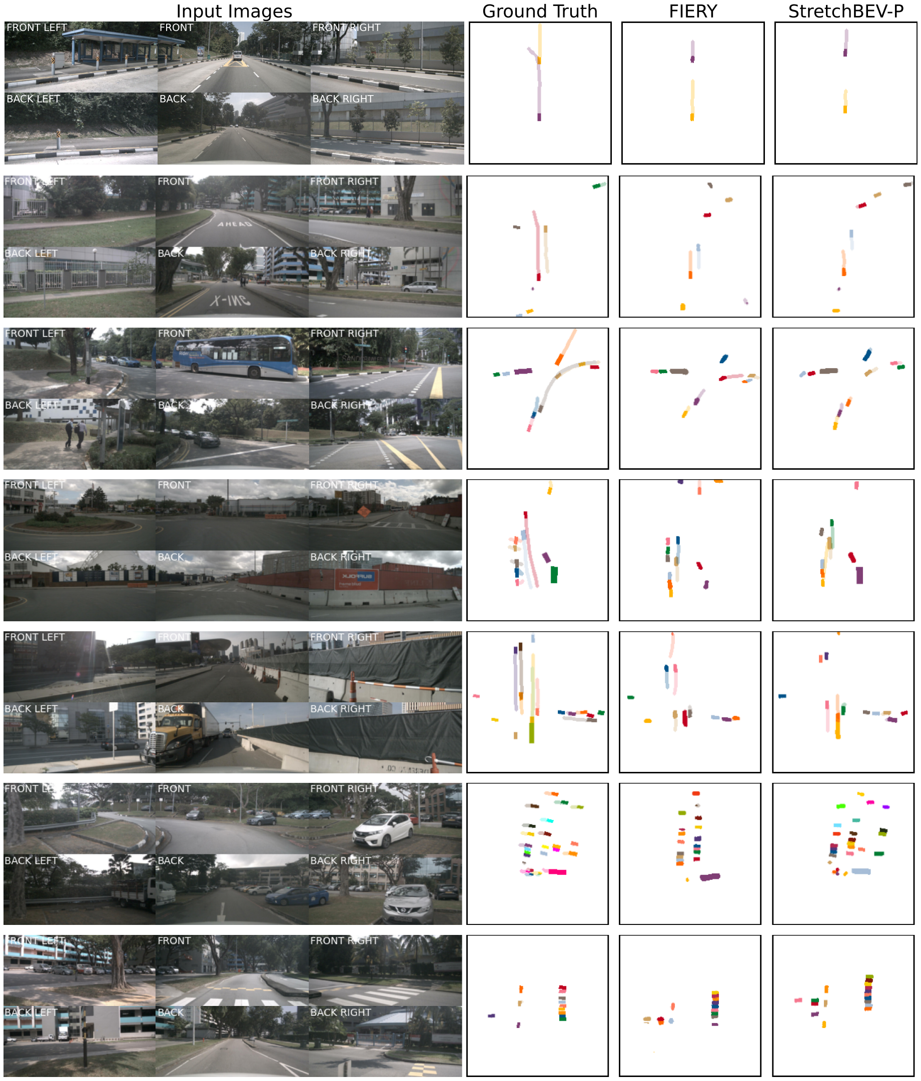

Fig. 3: **Qualitative Comparison for 6 seconds into future.** In this figure, we qualitatively compare the results of our model StretchBEV-P **(right)** to the ground truth **(left)** and FIERY [4] **(middle)** over long temporal horizon, which corresponds to predicting 6.0 seconds into the future. Each color represents an instance of a vehicle with its trajectory trailing in the same color transparently.

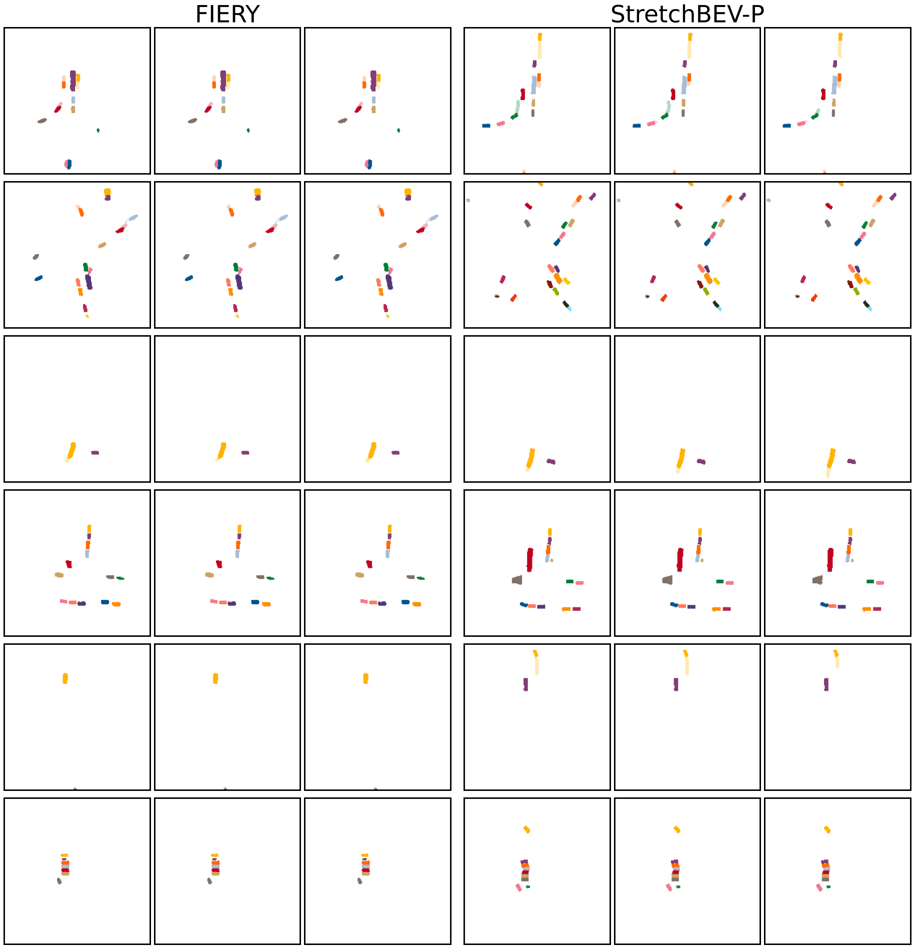

Fig. 4: **Qualitative Comparison of Diversity for** 2.0 **seconds into future.** In this figure, we visualize random samples from FIERY [4] **(left)** and our model StretchBEV-P **(right)** over short temporal horizon, which corresponds to predicting 2.0 seconds into future.

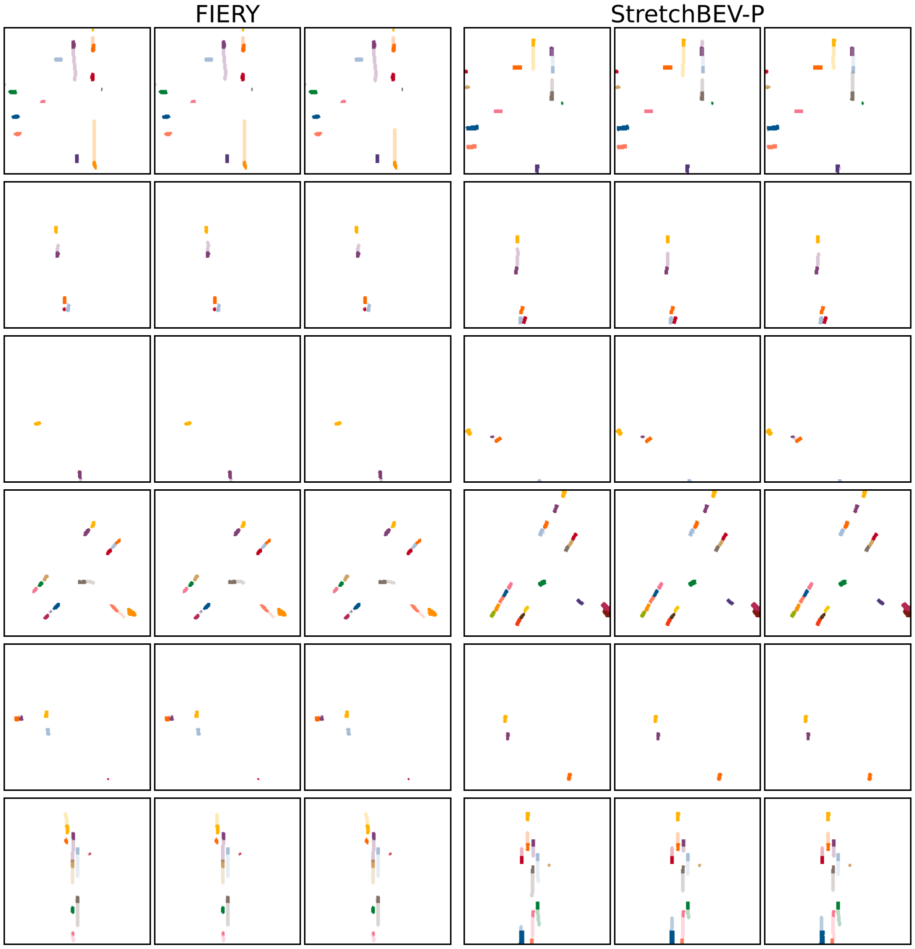

Fig. 5: **Qualitative Comparison of Diversity for** 4.0 **seconds into future.** In this figure, we visualize random samples from FIERY [4] **(left)** and our model StretchBEV-P **(right)** over mid temporal horizon, which corresponds to predicting 4.0 seconds into future.

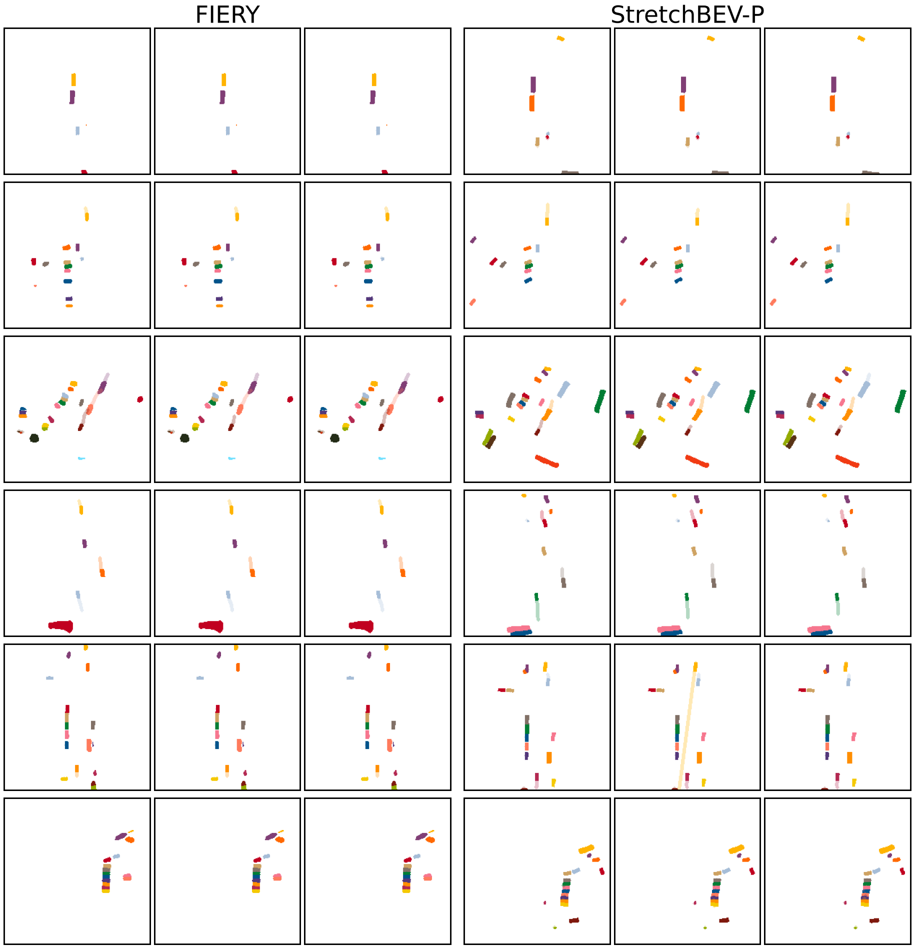

Fig. 6: **Qualitative Comparison of Diversity for** 6.0 **seconds into future.** In this figure, we visualize random samples from FIERY [4] **(left)** and our model StretchBEV-P **(right)** over long temporal horizon, which corresponds to predicting 6.0 seconds into future.

References

1. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2020)
2. Franceschi, J.Y., Delasalles, E., Chen, M., Lamprier, S., Gallinari, P.: Stochastic latent residual video prediction. In: Proc. of the International Conf. on Machine learning (ICML) (2020)
3. Hendy, N., Sloan, C., Tian, F., Duan, P., Charchut, N., Xie, Y., Wang, C., Philbin, J.: FISHING net: Future inference of semantic heatmaps in grids. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) Workshops (2020)
4. Hu, A., Murez, Z., Mohan, N., Dudas, S., Hawke, J., Badrinarayanan, V., Cipolla, R., Kendall, A.: FIERY: Future instance segmentation in bird's-eye view from surround monocular cameras. In: Proc. of the IEEE International Conf. on Computer Vision (ICCV) (2021)
5. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2018)
6. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Proc. of the International Conf. on Learning Representations (ICLR) (2014)
7. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: Proc. of the European Conf. on Computer Vision (ECCV) (2020)