Euro-PVI: Pedestrian Vehicle Interactions in Dense Urban Centers
– Supplemental Material –

Apratim Bhattacharyya[1] Daniel Olmeda Reino[2] Mario Fritz[3] Bernt Schiele[1]

[1]Max Planck Institute for Informatics, Saarland Informatics Campus
[2]Toyota Motor Europe
[3]CISPA Helmholtz Center for Information Security

1. Introduction

In the supplemental, we include 1. Further details about the Euro-PVI dataset (in Section 2). 2. Further details of our Joint-β-cVAE and additional evaluation (in Section 3).

In particular, in Section 2 we provide further details of the sensor setup, adding to the information in Section 3 of the main paper and also clarify the collection process of the Euro-PVI dataset. To further highlight the diversity of ego-vehicle - pedestrian (bicyclist) interactions in dense urban scenes, we include additional qualitative examples of the interactions in the Euro-PVI dataset in Fig. 8.

In Section 3, we elaborate the deviation of the ELBO for our Joint-β-cVAE approach (*c.f*. Eq. (5) in the main paper) and provide additional architectural details of our Joint-β-cVAE model (*c.f*. Sec. 5 in the main paper). Additionally, we also provide details of conditioning our Joint-β-cVAE model on visual features (*c.f*. Table 3 in the main paper), evaluation using ADE and FDE metrics with sample sizes of $N = \{3, 20\}$ (Tables 5 and 6), evaluation using the KDE NLL metric of models transferred from nuScenes [6] to Euro-PVI (Table 7) and evaluation of models trained both on nuScenes and Euro-PVI on nuScenes (Table 8). We also provide qualitative examples to further highlight the effectiveness of Joint-β-cVAE model in capturing ego-vehicle - pedestrian (bicyclist) interactions in Fig. 9.

2. Further Details about the Euro-PVI dataset

Further Details of Sensor Setup for Euro-PVI. The Euro-PVI dataset was recorded from a vehicle equipped with a lidar (Velodyne HDL-64E) mounted over the roof, which captures point clouds with a frequency of 10Hz. The vehicle also includes front-facing camera(s) installed behind the windshield which have a similar point of view as the driver. The cameras capture images at a resolution of 1280×806, and a frequency of 10Hz. Images are calibrated to remove distortion. This setup is adequate for capturing ego-vehicle - pedestrian (bicyclist) interactions as most interactions happen in front of the vehicle. Further, all sensors are registered to a common frame coordinate system inside the vehicle. Each data point has a corresponding position/pose from an on-board high performance GPS/IMU, and all are timestamped and synchronized. Such a setup also allows for the use of mapping services *e.g*. OpenStreetMap.

Examples of Interactions in the Euro-PVI dataset. We provide additional examples of interactions in the Euro-PVI dataset in Fig. 8 with five example interactions between the ego-vehicle and pedestrians (bicyclists) to highlight the diversity of interactions. Analogous to the Fig. 2 in the main paper, we also include the L_2 norms of the velocity and acceleration to illustrate the effect of interactions on the pedestrian (bicyclist) trajectories – which again highlights the need to model such ego-vehicle - pedestrian (bicyclist) interactions for accurate pedestrian (bicyclist) trajectory prediction.

3. Further Details of our Joint-β-CVAE and Additional Evaluation

We now provide additional details of our Joint-β-cVAE approach. We first detail the derivation of the ELBO (Eq. (5) in the main paper). Following this, we provide the details of the network architecture and the hyperparameters used in the Joint-β-cVAE model.

Details of the ELBO. From Eq. (2) in the main paper, the joint probability of the future trajectories of the agents in the scene can be expressed as,

$$p_\theta(\mathbf{Y}|\mathbf{X}) = \int \prod_i^n p_\theta(\mathbf{y}_i|\mathbf{Z}_{\leq i}, \mathbf{Y}_{<i}, \mathbf{X}) p_\theta(\mathbf{z}_i|\mathbf{Z}_{<i}, \mathbf{Y}_{<i}, \mathbf{X})\, d\mathbf{Z}. \tag{6}$$

Using the joint posterior as defined in Eq. (3) in the main

paper, the joint probability above can be expressed as,

$$p_\theta(\mathbf{Y}|\mathbf{X}) = \int \prod_i^n p_\theta(\mathbf{y}_i|\mathbf{Z}_{\leq i}, \mathbf{Y}_{<i}, \mathbf{X}) \frac{p_\theta(\mathbf{z}_i|\mathbf{Z}_{<i}, \mathbf{Y}_{<i}, \mathbf{X})}{q_\phi(\mathbf{z}_i|\mathbf{Z}_{<i}, \mathbf{X}, \mathbf{Y})} \quad (7)$$
$$q_\phi(\mathbf{z}_i|\mathbf{Z}_{<i}, \mathbf{X}, \mathbf{Y})\, d\mathbf{Z}.$$

Therefore, the log-likelihood of the joint distribution is,

$$\log(p_\theta(\mathbf{Y}|\mathbf{X})) = \log\left(\prod_i^n \mathbb{E}_{q_\phi}\Big(p_\theta(\mathbf{y}_i|\mathbf{Z}_{\leq i}, \mathbf{Y}_{<i}, \mathbf{X}) \frac{p_\theta(\mathbf{z}_i|\mathbf{Z}_{<i}, \mathbf{Y}_{<i}, \mathbf{X})}{q_\phi(\mathbf{z}_i|\mathbf{Z}_{<i}, \mathbf{X}, \mathbf{Y})}\Big)\right) \quad (8)$$
$$= \sum_i^n \log\left(\mathbb{E}_{q_\phi}\Big(p_\theta(\mathbf{y}_i|\mathbf{Z}_{\leq i}, \mathbf{Y}_{<i}, \mathbf{X}) \frac{p_\theta(\mathbf{z}_i|\mathbf{Z}_{<i}, \mathbf{Y}_{<i}, \mathbf{X})}{q_\phi(\mathbf{z}_i|\mathbf{Z}_{<i}, \mathbf{X}, \mathbf{Y})}\Big)\right).$$

Now, using Jensen's inequality and introducing the β term to weigh the KL-divergence term (D_{KL}) as in [17] gives us the ELBO in Eq. (5) of the main paper,

$$\log(p_\theta(\mathbf{Y}|\mathbf{X})) \geq \sum_i \mathbb{E}_{q_\phi} \log(p_\theta\big(\mathbf{y}_i|\mathbf{Z}_{\leq i}, \mathbf{Y}_{<i}, \mathbf{X}\big)) - \beta \sum_i D_{\text{KL}}\big(q_\phi(\mathbf{z}_i|\mathbf{Z}_{<i}, \mathbf{X}, \mathbf{Y})||p_\theta(\mathbf{z}_i|\mathbf{Z}_{<i}, \mathbf{X})\big).$$

Additional Architectural Details. We model the posterior distribution (q_ϕ) using LSTMs with 128 hidden neurons. The attention over each of the previously sampled $\mathbf{z}_j \in \mathbf{Z}_{<i}$ and $\mathbf{x}_j \in \mathbf{X}, \mathbf{y}_j \in \mathbf{Y}$ is modeled using fully connected layers with 64 hidden units. Similarly, the prior, p_θ is modeled using an LSTM with 128 hidden units and the attention over each of the previously sampled $\mathbf{z}_j \in \mathbf{Z}_{<i}$ and $\mathbf{x}_j \in \mathbf{X}$ is modeled using fully connected layers with 64 hidden units. The decoder is also modeled using an LSTM with 128 hidden units. We use a latent space of 32 dimensions. We find that β value of $[0.08, 0.12]$ works well in practice and helps us learn representative latent spaces. We use the Adam [46] optimizer with a learning rate of 3×10^{-3} with a exponential decay of 0.9999.

Details of Conditioning on Visual Features. In Table 3 in the main paper, we report results on our Euro-PVI dataset where our Joint-β-cVAE model is additionally conditioned on visual features. We use both the RGB camera images and lidar point clouds. In detail, we use a 256×256 crop of the camera image and a 5 meter $\times$ 5 meter bird eye view rendering of the lidar point cloud both centered at the pedestrian (bicyclist). The latent space of our Joint-β-cVAE is additionally conditioned on these visual features using a simple VGG-16 [47] like neural network. We see that the performance of our Joint-β-cVAE further improves, because the camera image and bird eye view lidar provides important contextual information *e.g.* physical obstacles in the vicinity of the pedestrian (bicyclist) which can have a significant impact on the trajectory of the pedestrian (bicyclist).

Additional Metrics on Euro-PVI. Here, we additionally report for Trajectron++ [37] and our Joint-β-cVAE, the average (euclidean) displacement error (ADE) at $t + \{1, 2, 3\}$ seconds for $N = 20$ samples in Table 5 and both the final (euclidean) displacement error (FDE, equivalent to the Best of N error in Table 2,3 of the main paper) and average (euclidean) displacement error (ADE) at $t + \{1, 2, 3\}$ seconds for $N = 3$ samples in Table 6. The results are consistent with the findings in Table 2,3 of the main paper where our Joint-β-cVAE outperforms Trajectron++ [37] for both sample sizes $N = \{3, 20\}$.

Additional Metrics on Transferring from nuScenes to Euro-PVI. While in Table 4 (in the main paper), we report only the Best of N metric, in Table 7, we additionally report the KDE NLL metric. Similar to the observations with the Best of N metric, observe a considerable drop in performance as measured by the KDE NLL metric in comparison to the performance of the models when they are both trained and evaluated on Euro-PVI. Again, this provides additional evidence that the distribution of trajectories and interaction patterns in Euro-PVI is significantly different compared to nuScenes.

Transferring from Euro-PVI to nuScenes. In Table 4 (in the main paper), we show that models trained (only) on nuScenes [6] do not perform well when evaluated on Euro-PVI. On the other hand, in Table 8, we show that training both on Euro-PVI and nuScenes can improve performance on nuScenes. In particular, we consider the more challenging setting we provide a shorter observation of 1 second to our Joint-β-cVAE. The performance on nuScenes improves because Euro-PVI also contains significant pedestrian (bicyclist) - pedestrian (bicyclist) interactions in addition to vehicle - pedestrian (bicyclist) interactions.

Further Qualitative Examples. To further validate our Joint-β-cVAE approach for modeling complex interactions in dense urban scenarios, we provide additional qualitative examples of predictions on the Euro-PVI dataset in Fig. 9. We compare the predictions of our Joint-β-cVAE approach to that of Trajectron++ [37] using the Best of $N = 20$ samples. In Fig. 9 (top left) we see that our Joint-β-cVAE model correctly predicts that the pedestrian stays on the sidewalk and waits for the ego-vehicle to pass. Similarly, in Fig. 9 (top right) our Joint-β-cVAE model correctly predicts that the pedestrian crosses the street in front of the ego-vehicle. In Fig. 9 (middle left) our Joint-β-cVAE model correctly predicts that the bicyclist yields to the on-coming ego-vehicle. In Fig. 9 (middle right) our Joint-β-cVAE model correctly predicts that the bicyclist continues straight ahead and does not attempt to cross the street due to the on-coming ego-

Method	Interactions		FDE $N=20\downarrow$			ADE $N=20\downarrow$		
	P-P	**P-V**	$t+1$ **sec**	$t+2$ **sec**	$t+3$ **sec**	$t+1$ **sec**	$t+2$ **sec**	$t+3$ **sec**
Trajectron++ [37]	✓	✓	**0.09**	0.28	0.54	**0.05**	0.13	0.24
Joint-β-cVAE (Ours)	✓	✓	**0.09**	**0.27**	**0.51**	**0.05**	**0.12**	**0.23**

Table 5. Additional metrics on Euro-PVI with $N=20$ samples. P-P and P-V: whether pedestrian - pedestrian or pedestrian - ego-vehicle interactions are modeled.

Method	Interactions		FDE $N=3\downarrow$			ADE $N=3\downarrow$		
	P-P	**P-V**	$t+1$ **sec**	$t+2$ **sec**	$t+3$ **sec**	$t+1$ **sec**	$t+2$ **sec**	$t+3$ **sec**
Trajectron++ [37]	✓	✓	0.18	0.53	1.01	**0.09**	0.22	0.42
Joint-β-cVAE (Ours)	✓	✓	**0.17**	**0.51**	**0.99**	**0.09**	**0.22**	**0.41**

Table 6. Additional metrics on Euro-PVI with $N=3$ samples. P-P and P-V: whether pedestrian - pedestrian or pedestrian - ego-vehicle interactions are modeled.

Method	Interactions		Best of $N=20\downarrow$			KDE NLL $\downarrow$		
	P-P	**P-V**	$t+1$ **sec**	$t+2$ **sec**	$t+3$ **sec**	$t+1$ **sec**	$t+2$ **sec**	$t+3$ **sec**
Trajectron++ [37]	✓	–	0.10	0.35	0.63	-1.07	0.15	1.45
Trajectron++ [37]	✓	✓	0.10	0.35	0.63	-1.04	0.14	1.42
Joint-β-cVAE (Ours)	✓	–	0.10	0.33	0.60	-1.51	-0.09	1.22
Joint-β-cVAE (Ours)	✓	✓	0.10	0.33	0.61	-1.56	-0.10	1.31

Table 7. Transferring models trained on nuScenes to Euro-PVI (see also Table 4 in the main paper).

Method	Training		Best of $N=20\downarrow$			KDE NLL $\downarrow$		
	nuScenes	**Euro-PVI**	$t+1$ **sec**	$t+2$ **sec**	$t+3$ **sec**	$t+1$ **sec**	$t+2$ **sec**	$t+3$ **sec**
Joint-β-cVAE (Ours)	✓	–	0.01	0.14	0.31	-0.20	1.97	3.56
Joint-β-cVAE (Ours)	✓	✓	0.01	0.13	0.30	-0.27	1.64	2.92

Table 8. Evaluating models trained on both on nuScenes and Euro-PVI on nuScenes (only pedestrian - pedestrian interactions are modeled).

vehicle. In Fig. 9 (bottom left) our Joint-β-cVAE model correctly predicts that the pedestrian crosses the street and the ego-vehicle yields to the pedestrian. Similarly, in Fig. 9 (bottom right) our Joint-β-cVAE model correctly predicts that both pedestrians cross the street (while avoiding collisions) and the the ego-vehicle yields to the pedestrians. These examples show that our Joint-β-cVAE model can successfully capture the effect of interactions on the multi-modal distribution of pedestrian trajectories in the latent space.

References

[46] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

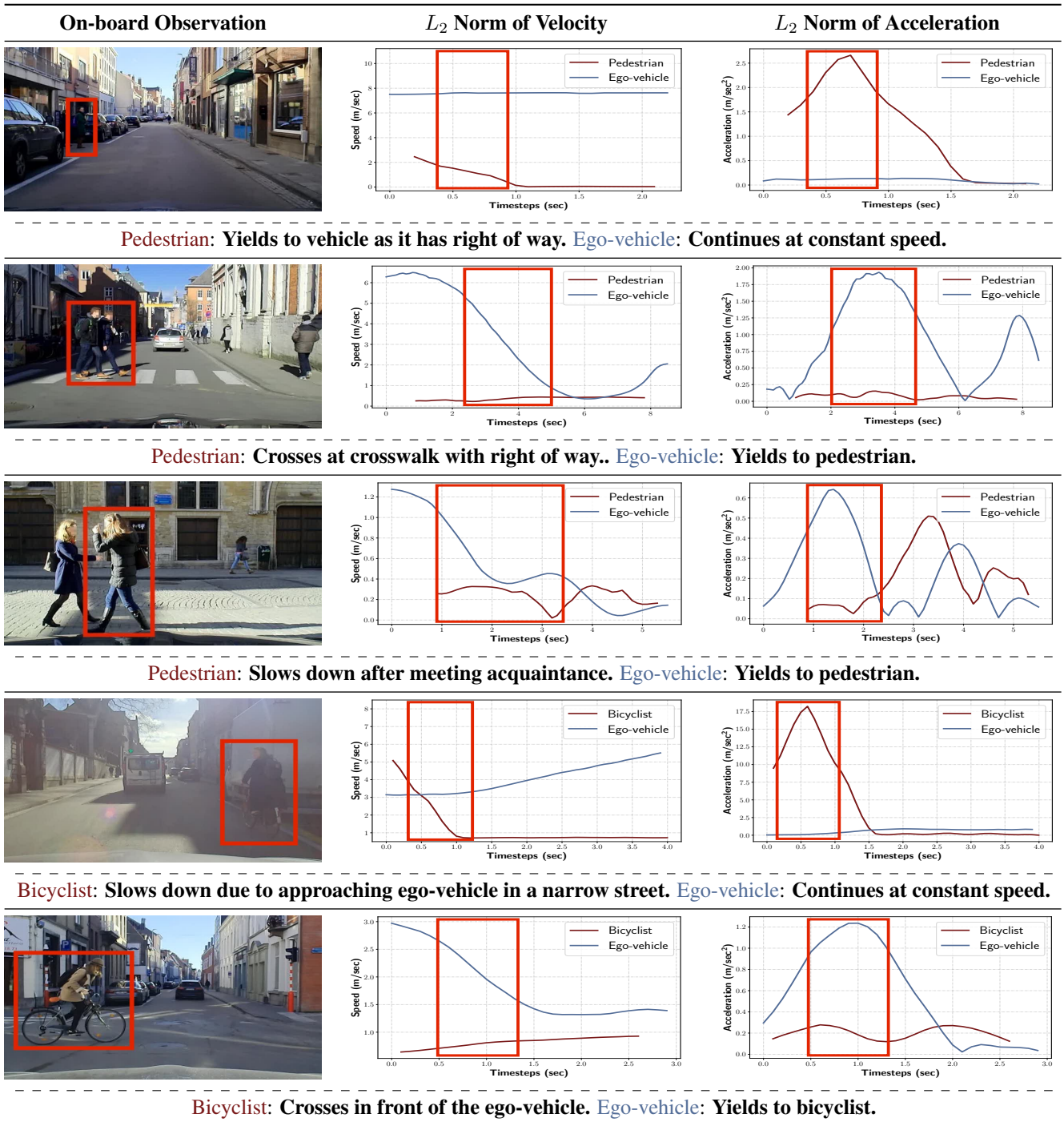

Figure 8. Examples of interactions in the proposed Euro-PVI dataset. Spikes in the magnitude (L_2 norm) of acceleration resulting from interactions are marked. Top row: the pedestrian attempting to cross the road yields to the on-coming ego-vehicle as it has right of way. Second row: the ego-vehicle yields to the pedestrians as they are at a crosswalk and thus have the right of way. Third row: the pedestrian crossing the street in front of the ego-vehicle slows down after meeting an acquaintance and the ego-vehicle yields to the pedestrians and waits for them to cross the street. Fourth row: the bicyclist slows down to let the ego-vehicle pass due to the narrow street. Fifth row: the bicyclist crosses in front of the ego-vehicle and the ego-vehicle yields to the bicyclist.

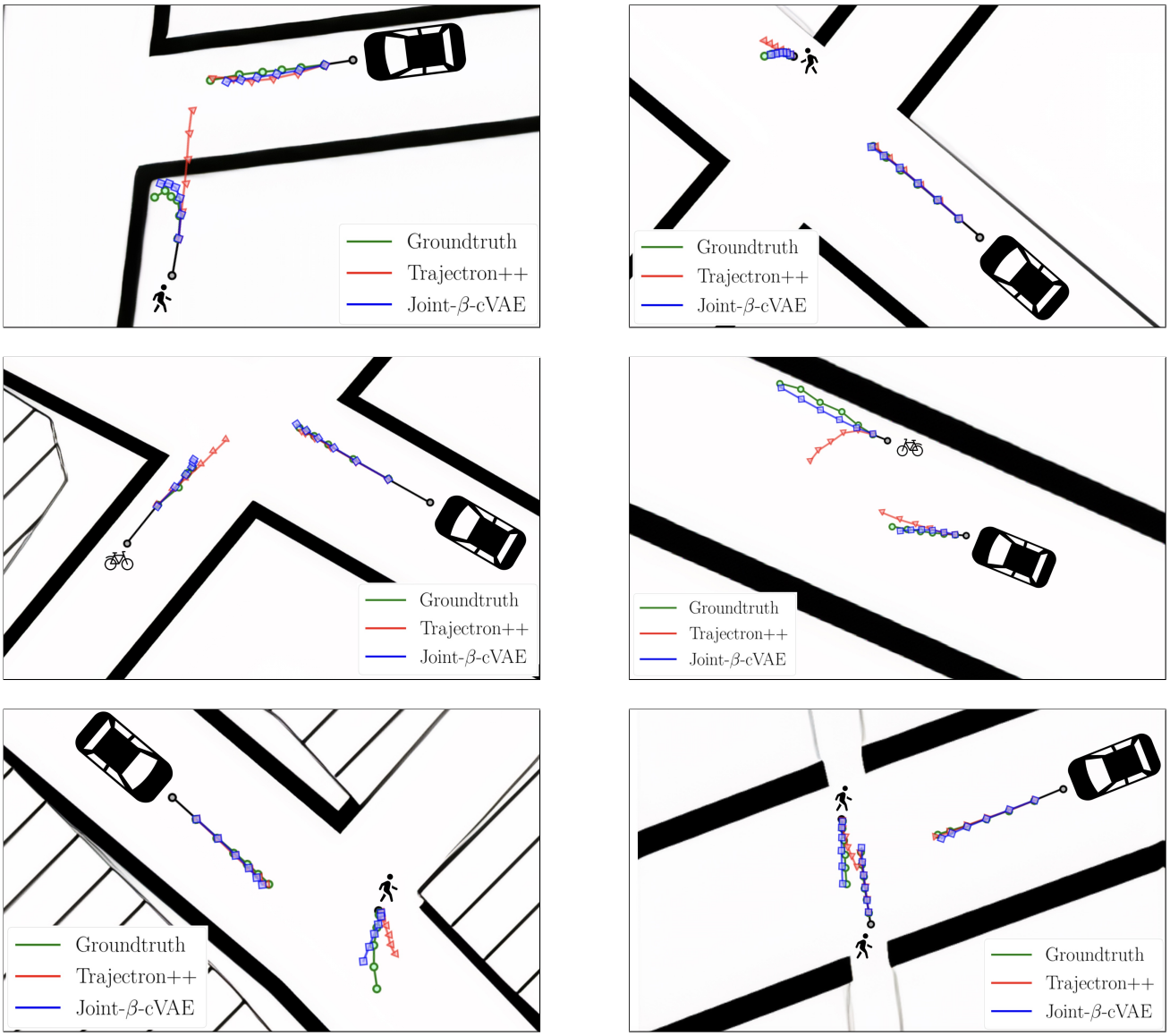

Figure 9. Qualitative examples on the Euro-PVI dataset. We compare the Best of $N = 20$ samples for Trajectron++ (red) and our Joint-β-cVAE (blue).