1 Supplementary

In the supplementary, we include some visualization results and details about how the communication volume is calculated.

1.1 Visualization

In Figure 1 and 2 we visualize some examples from the test set, including the ground truth multi-view observed scene, the completed scene and the results on detection and segmentation respectively. In addition, we also visualize a few examples together with the difference between the true observation and the completed scene in Figure 3 to give a clearer look at the completion quality.

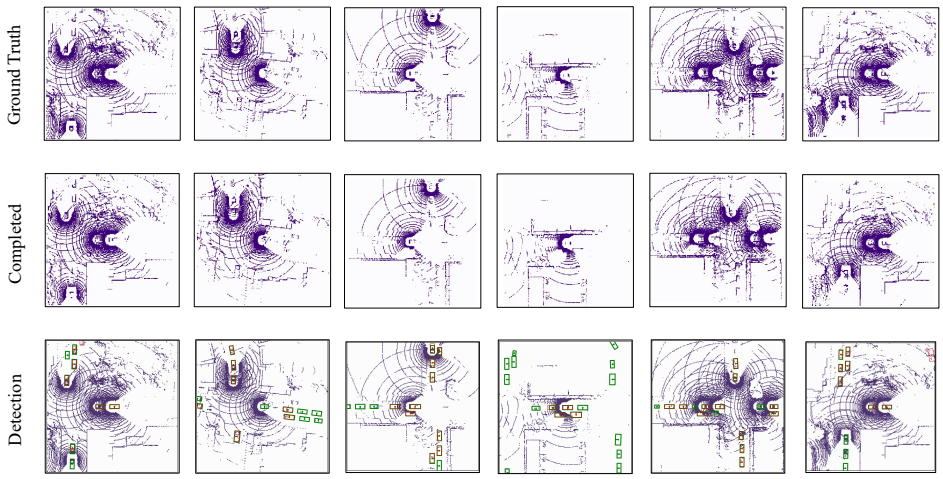

Figure 1: **Visualization: completion and detection**. 6 randomly sampled examples in the test set are visualized above. Rows from top to bottom are respectively: the ground truth multi-view scene, the completed scene predicted and the detection results given by the detection model based on the completed scene.

1.2 Communication Volume

In this work, we measure bandwidth to compare the communication volume required by different methods. Here we present the details of how it is calculated. In a nutshell, the robots communicate with intermediate representations, so to measure the bandwidth is to measure the size of the intermediate features being transmitted per second between robots. Specifically, if the intermediate feature has size $h \times w \times c$ and the model transmit $p\%$ following the time amortized approach, then the *byte size* of the data being transmitted per sample will be: $8 \times p\% \times h \times w \times c$ since each element of the feature is 8-byte floating point number. This can be generalized to other data types as well. Then assume the robot observe and communicate at a frequency of f ($f = 5$Hz for the V2X-Sim dataset), the communication bandwidth V_c is computed as:

$$V_c = f \times 8 \times p\% \times h \times w \times c \quad \text{Byte/s} \tag{1}$$

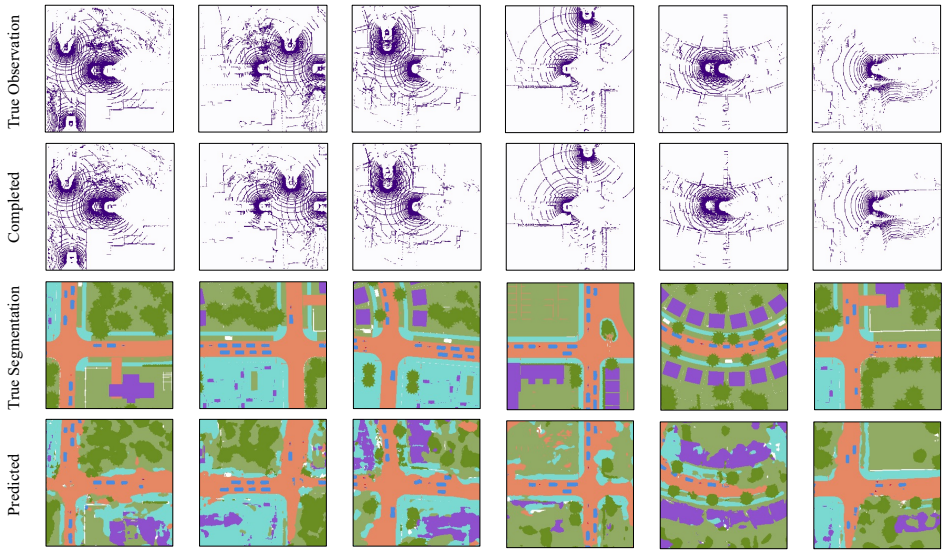

Figure 2: **Visualization: completion and segmentation**. 6 randomly sampled examples in the test set are visualized above. Rows from top to bottom are respectively: the ground truth multi-view scene, the completed scene, the ground truth semantic segmentation and the predicted results based on the completed scene.

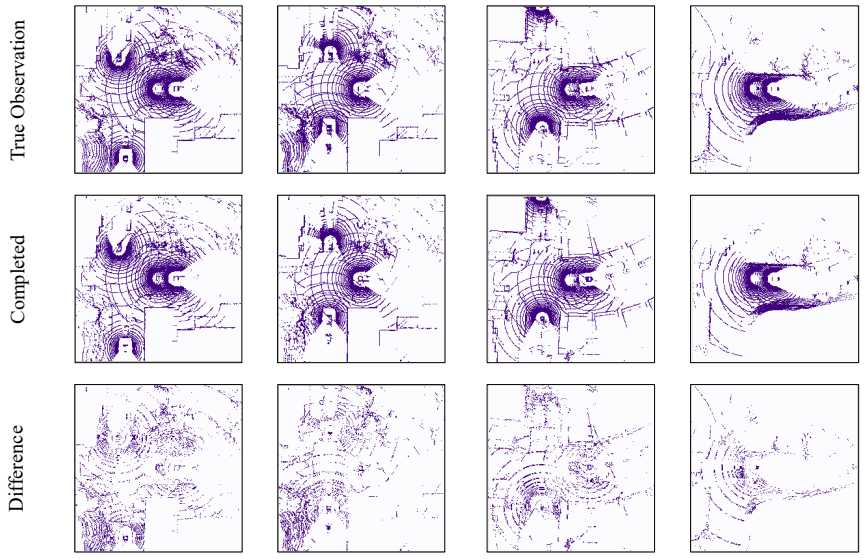

Figure 3: **Visualization: completion quality**. The first two rows are the true observation and the completed scene, and the last row shows the difference between them.