# DONUT: A Decoder-Only Model for Trajectory Prediction

Supplementary Material

## 1. LineAttention

In preliminary experiments, we found it beneficial to give the agent better access to road elements. Our baseline, QCNet, uses the beginning of a road polyline as the reference point for relative positional encodings. In addition to this, we add an encoding for relative information between the agent and its closest point on each map polyline, which we call *LineAttention*. However, for the final model, this procedure only had a tiny effect, decreasing minFDE from 1.181 to 1.176.

## 2. Efficiency Analysis

We measure inference time and show the number of parameters in Tab. 1. Due to the temporal unrolling, switching from the baseline to decoder-only almost triples the inference time. The refinement layer roughly doubles the number of successive operations and thus also the inference time. Training times behave similarly. Overprediction has negligible impact on efficiency and is dropped for inference. Note, however, that we did not focus on optimizing the code for efficiency, but instead on improving the prediction accuracy.

| DONUT | Ref. | Inference time (ms) | Num. parameters |
|---|---|---|---|
| ✗ | N/A | 23.7 | 7.7M |
| ✓ | ✗ | 65.7 | 5.2M |
| ✓ | ✓ | 129.0 | 9.0M |

Table 1. **Efficiency analysis** on an Nvidia RTX 4090 GPU.

## 3. Difficult Scenes

To assess DONUT on more challenging scenarios, we evaluate it on Argoverse 2 trajectories with a ground-truth future turn of at least 45° in Tab. 2. The relative improvement with respect to the encoder-decoder baseline becomes notably larger than on the full dataset (14.6% vs. 6.1% minFDE), showing that DONUT's periodic updates are especially helpful in complex situations.

| DONUT | Overp. | Ref. | b-minFDE | minFDE | minADE | MR |
|---|---|---|---|---|---|---|
| ✗ | N/A | N/A | 3.008 | 2.394 | 1.176 | 0.362 |
| ✓ | ✗ | ✗ | 2.766 | 2.129 | 1.147 | 0.308 |
| ✓ | ✓ | ✗ | 2.725 | 2.078 | 1.109 | 0.308 |
| ✓ | ✗ | ✓ | 2.757 | 2.148 | 1.160 | 0.329 |
| ✓ | ✓ | ✓ | **2.672** | **2.043** | **1.092** | **0.295** |

Table 2. **Results on Argoverse 2, only considering turns** $> 45°$**.**

## 4. Tokenizer Details

In Fig. 1 we visualize our tokenizer's architecture in detail. The 8-dimensional features for each time step consist of position and heading relative to the reference point, motion vectors, angular motion, velocity, and the difference of the heading and the motion vector direction. Type embeddings describe the object types (*e.g.*, car, bus, pedestrian) present in Argoverse 2.

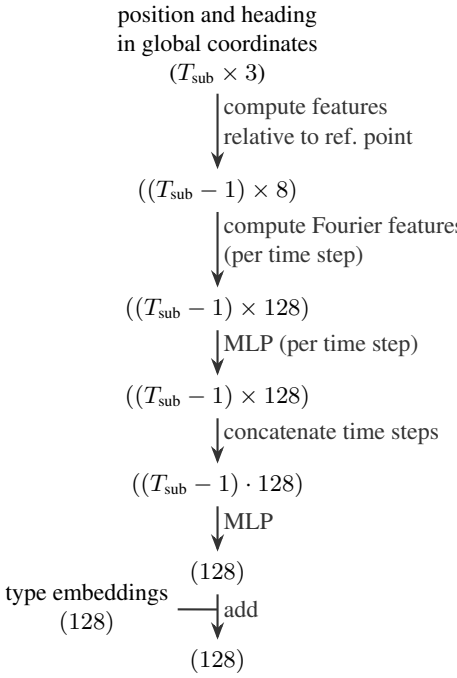



Figure 1. **Detailed tokenizer architecture.**

## 5. Failure Cases

We manually examined 100 scenes with a minFDE $> 5\,\text{m}$. Most errors are caused by predictions being too slow (27%) or too fast (19%), or missing a turn (19%). Additionally, 27% had rare ground-truth events, *e.g.*, vehicles moving off the road or maneuvering illegally. We visualize a few scenes in Fig. 2.

## 6. Additional Qualitative Results

We provide additional non-cherrypicked qualitative results in Figs. 3 to 18.

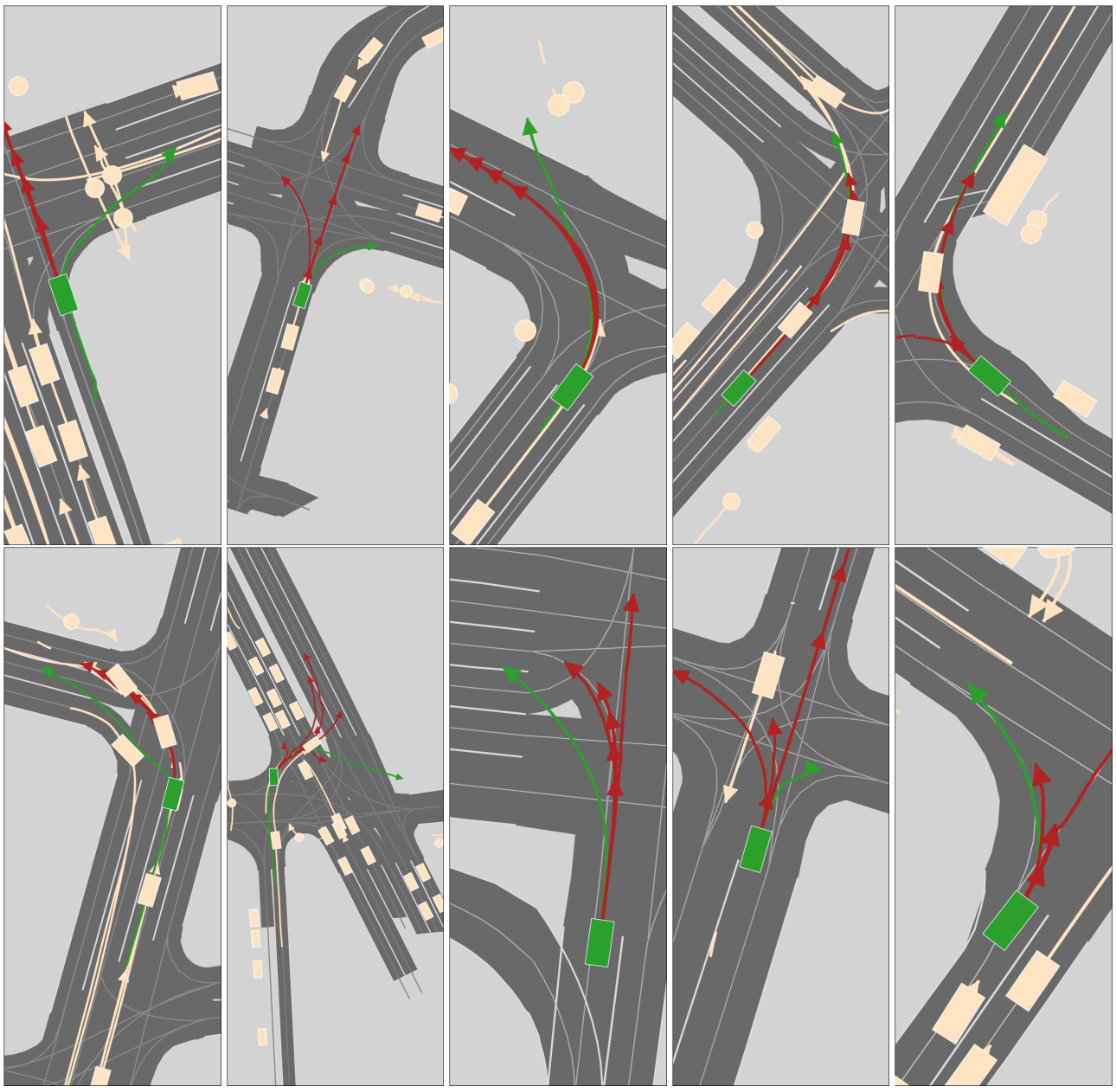
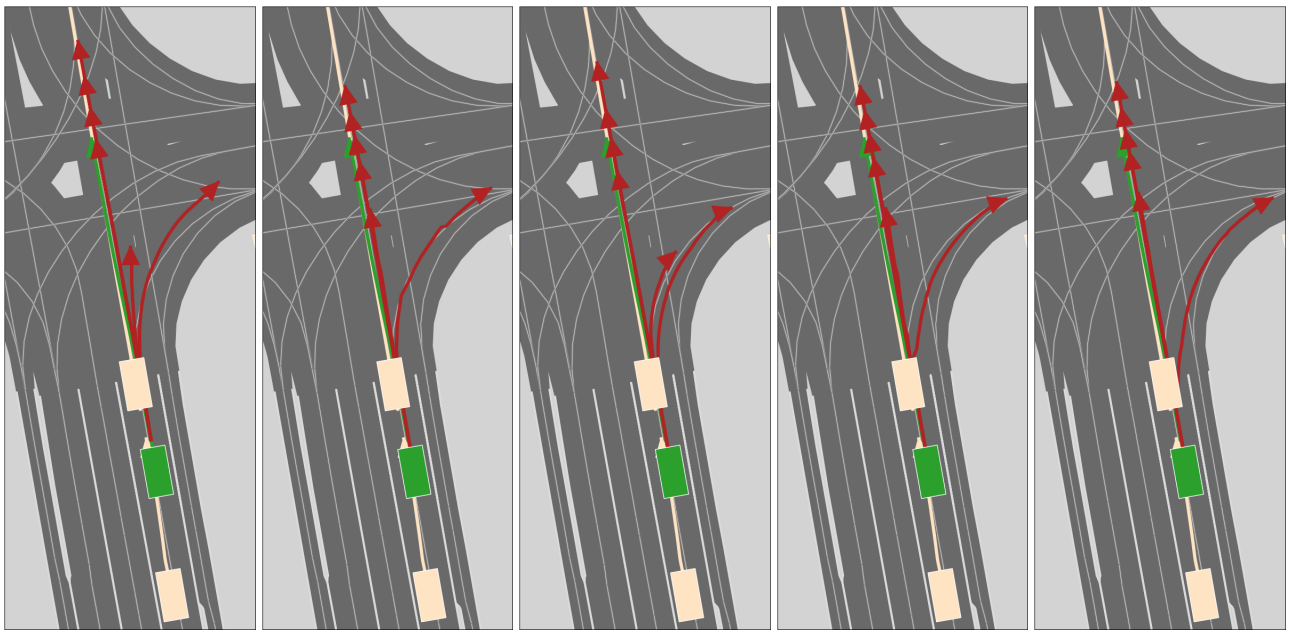
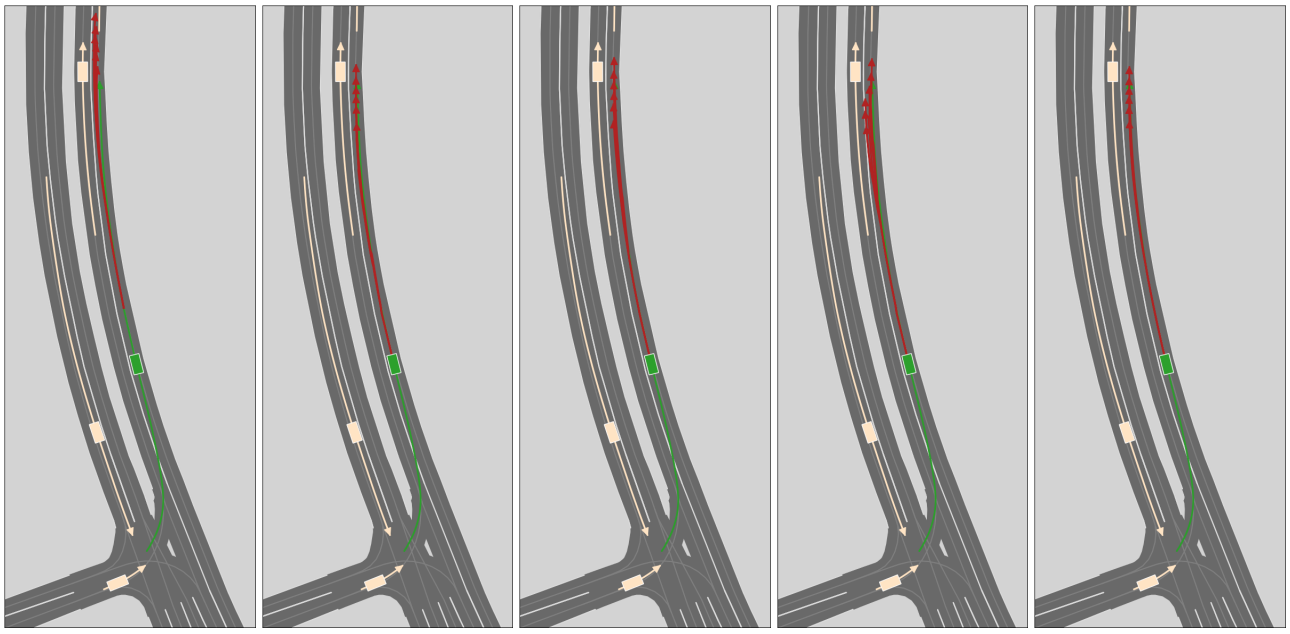


Figure 2. **Failure cases of DONUT.**



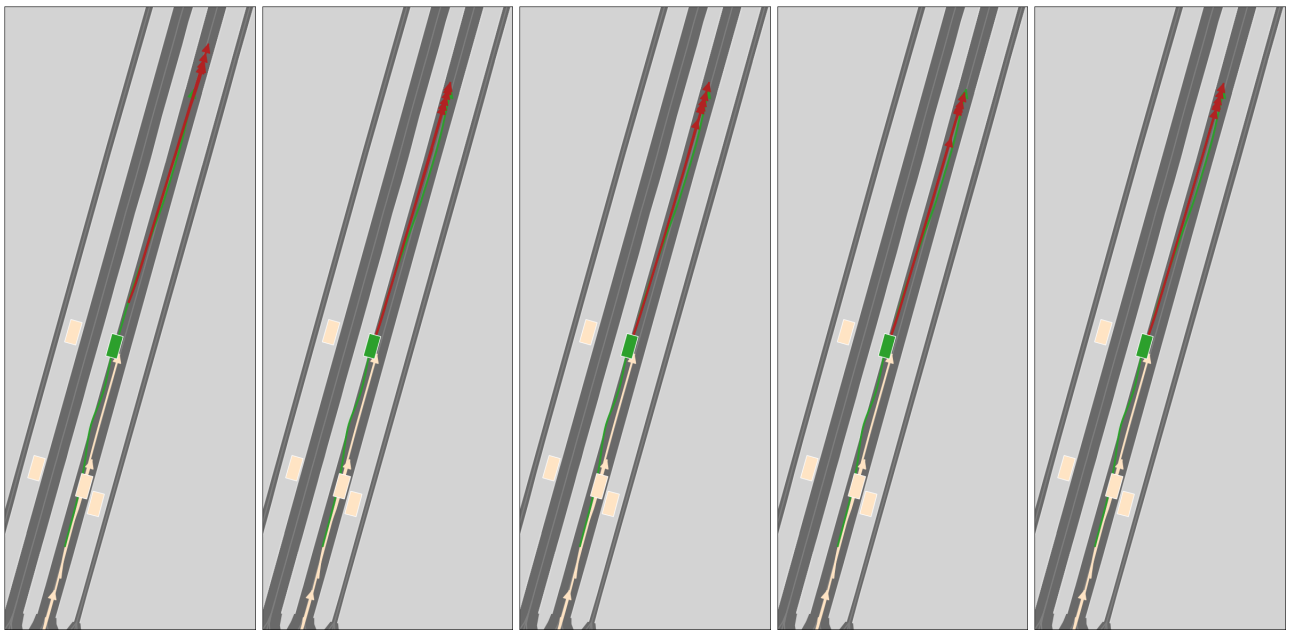
(a) Encoder-decoder (b) DONUT Overp. ✗ Ref. ✗ (c) DONUT Overp. ✓ Ref. ✗ (d) DONUT Overp. ✗ Ref. ✓ (e) DONUT Overp. ✓ Ref. ✓

Figure 3. **Additional qualitative results.**



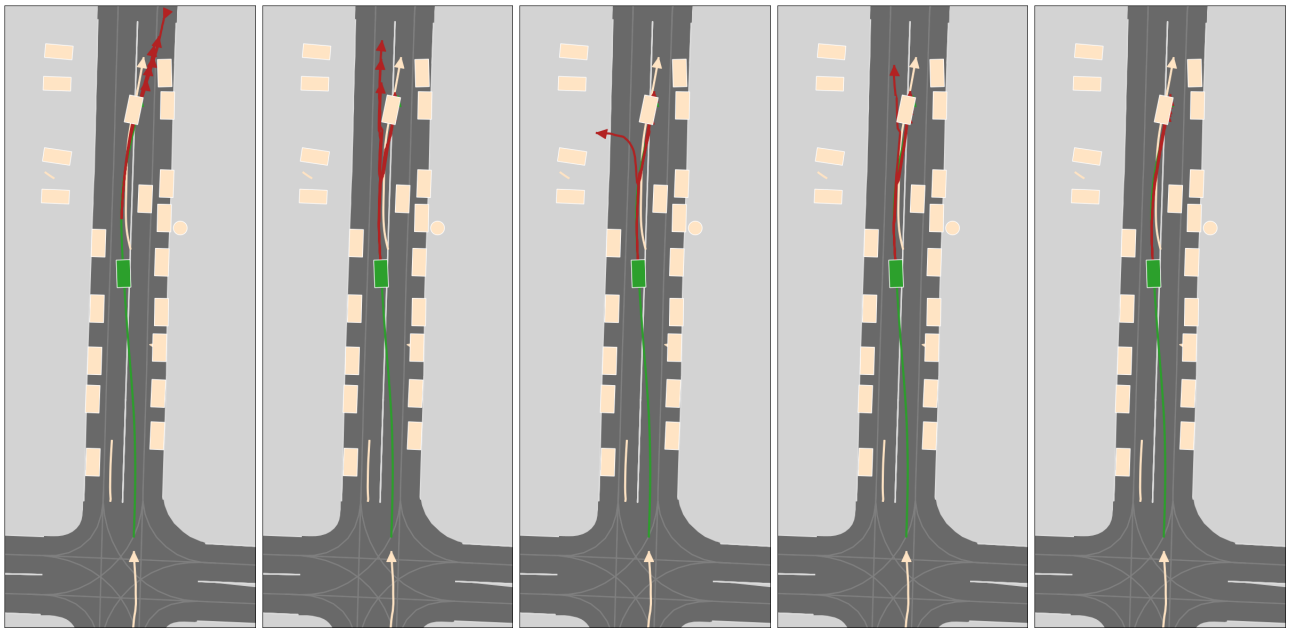
(a) Encoder-decoder (b) DONUT Overp. ✗ Ref. ✗ (c) DONUT Overp. ✓ Ref. ✗ (d) DONUT Overp. ✗ Ref. ✓ (e) DONUT Overp. ✓ Ref. ✓

Figure 4. **Additional qualitative results.**



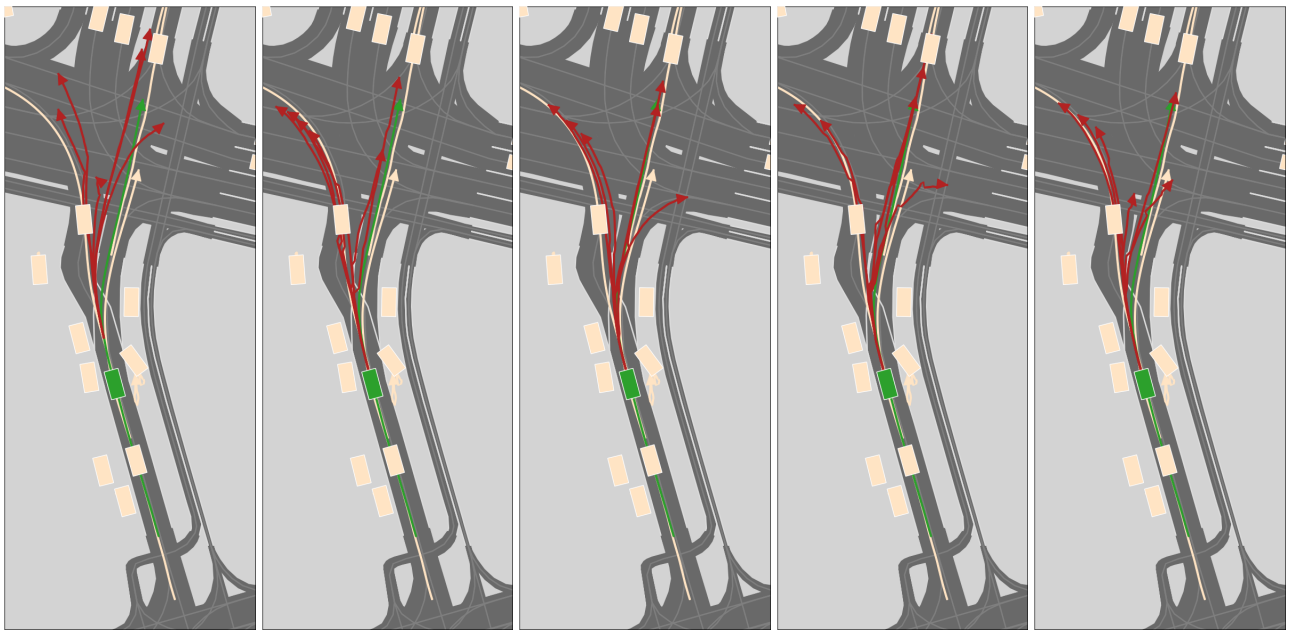
(a) Encoder-decoder (b) DONUT Overp. ✗ Ref. ✗ (c) DONUT Overp. ✓ Ref. ✗ (d) DONUT Overp. ✗ Ref. ✓ (e) DONUT Overp. ✓ Ref. ✓

Figure 5. **Additional qualitative results.**



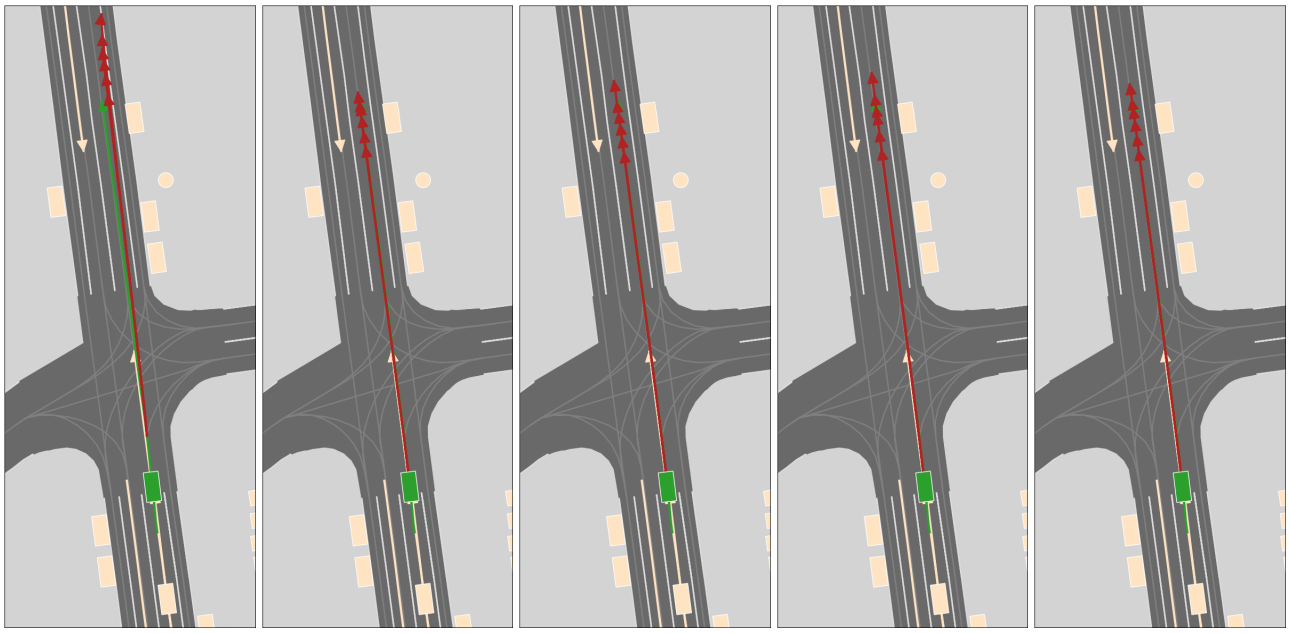
(a) Encoder-decoder (b) DONUT Overp. ✗ Ref. ✗ (c) DONUT Overp. ✓ Ref. ✗ (d) DONUT Overp. ✗ Ref. ✓ (e) DONUT Overp. ✓ Ref. ✓

Figure 6. **Additional qualitative results.**



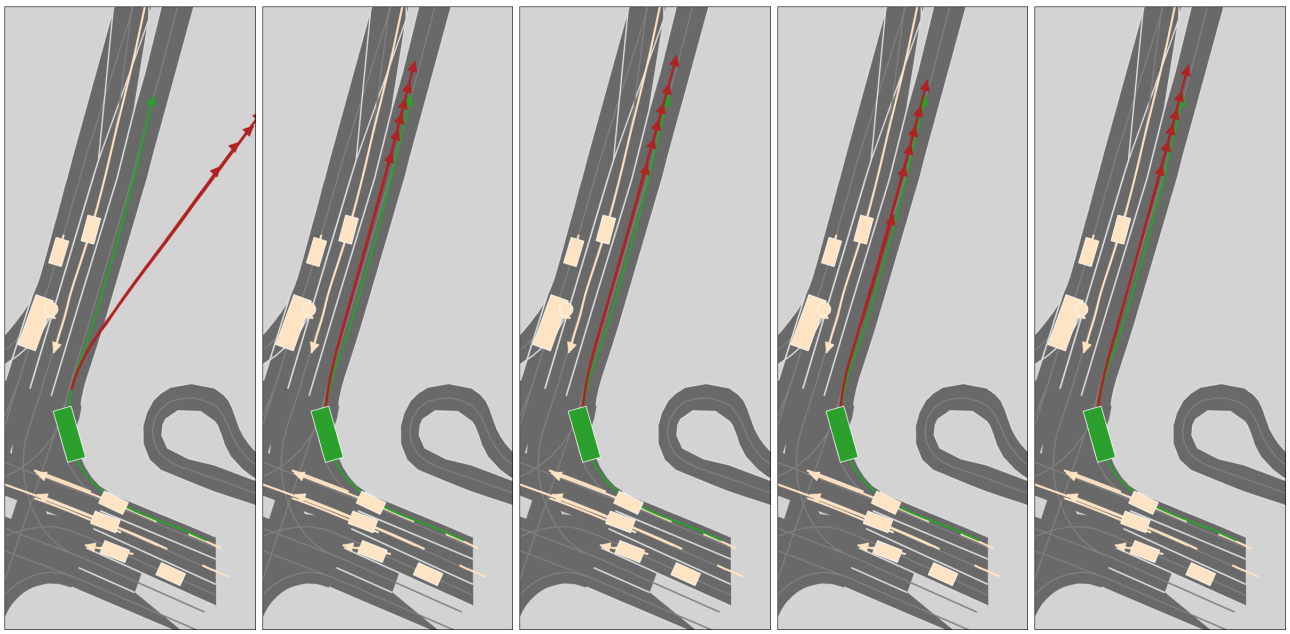
(a) Encoder-decoder (b) DONUT Overp. ✗ Ref. ✗ (c) DONUT Overp. ✓ Ref. ✗ (d) DONUT Overp. ✗ Ref. ✓ (e) DONUT Overp. ✓ Ref. ✓

Figure 7. **Additional qualitative results.**



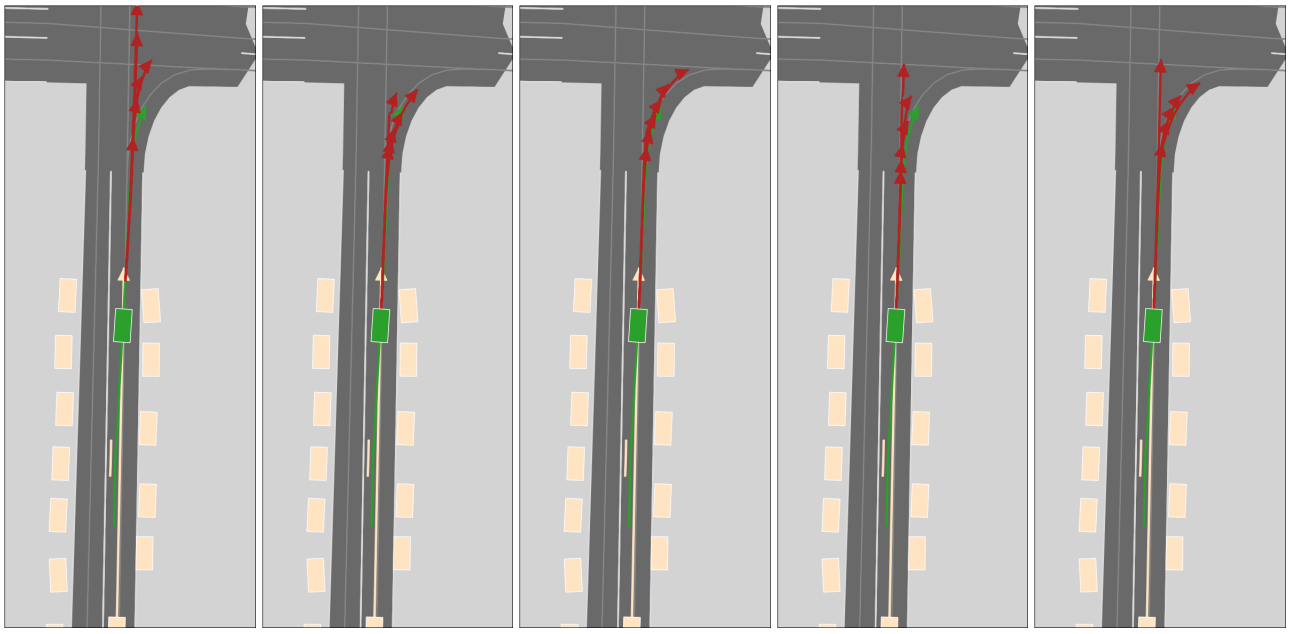
(a) Encoder-decoder (b) DONUT Overp. ✗ Ref. ✗ (c) DONUT Overp. ✓ Ref. ✗ (d) DONUT Overp. ✗ Ref. ✓ (e) DONUT Overp. ✓ Ref. ✓

Figure 8. **Additional qualitative results.**



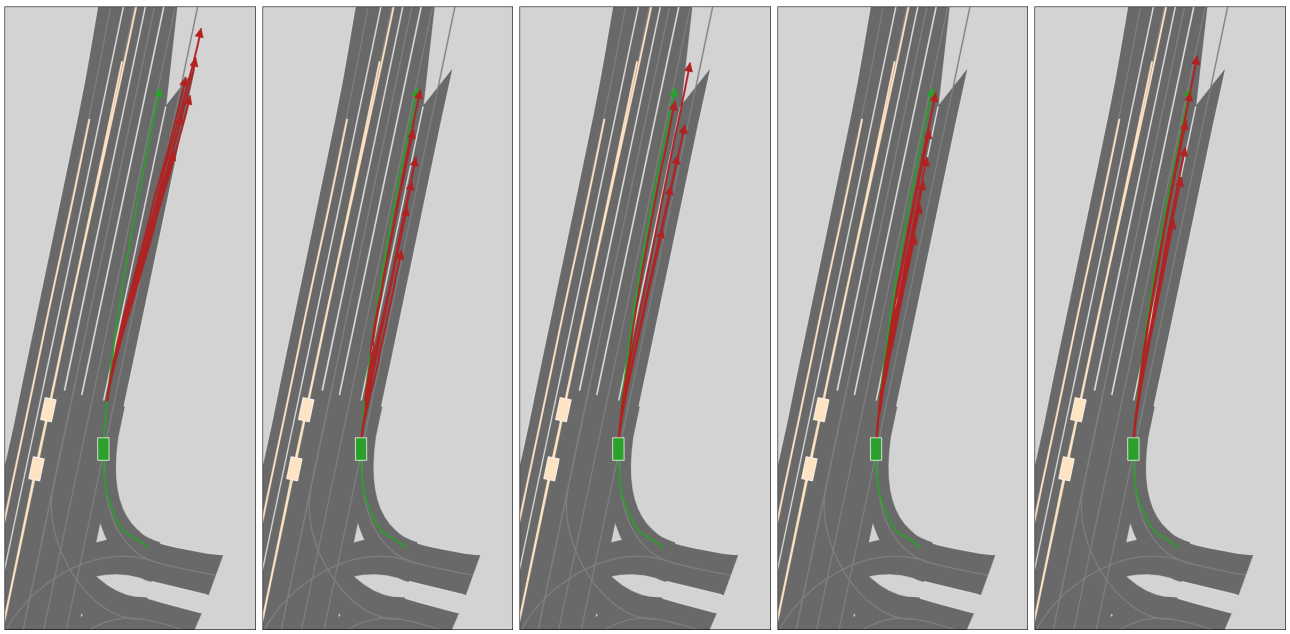
(a) Encoder-decoder (b) DONUT Overp. ✗ Ref. ✗ (c) DONUT Overp. ✓ Ref. ✗ (d) DONUT Overp. ✗ Ref. ✓ (e) DONUT Overp. ✓ Ref. ✓

Figure 9. **Additional qualitative results.**



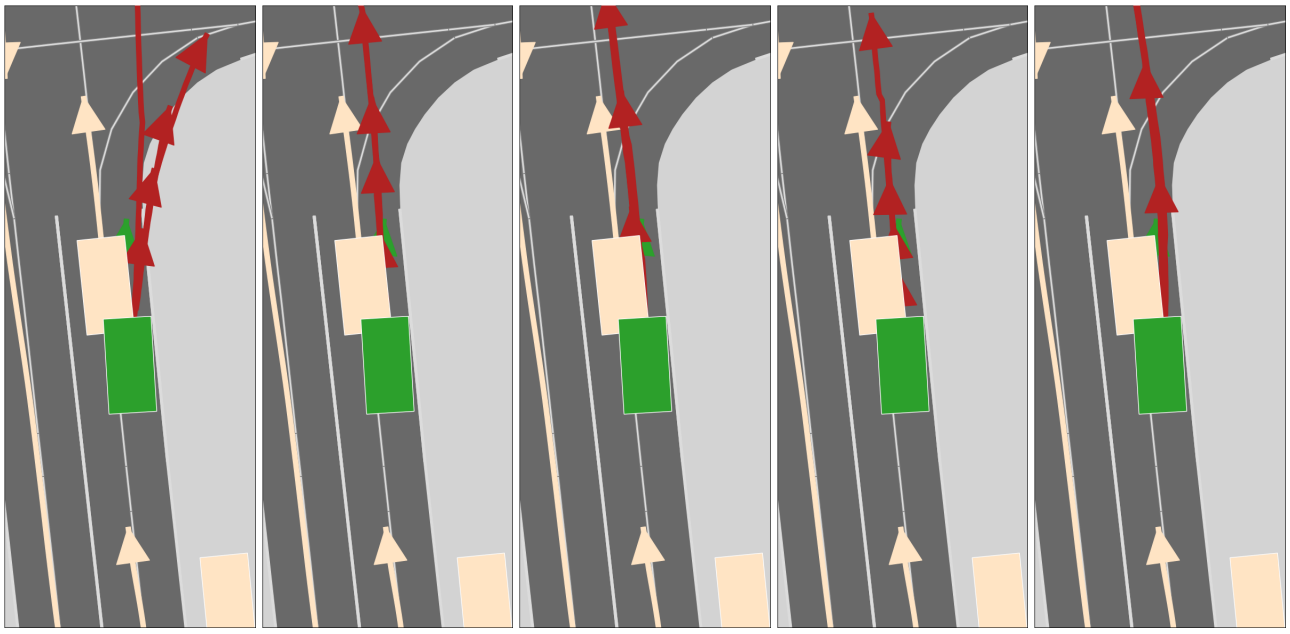
(a) Encoder-decoder (b) DONUT Overp. ✗ Ref. ✗ (c) DONUT Overp. ✓ Ref. ✗ (d) DONUT Overp. ✗ Ref. ✓ (e) DONUT Overp. ✓ Ref. ✓

Figure 10. **Additional qualitative results.**



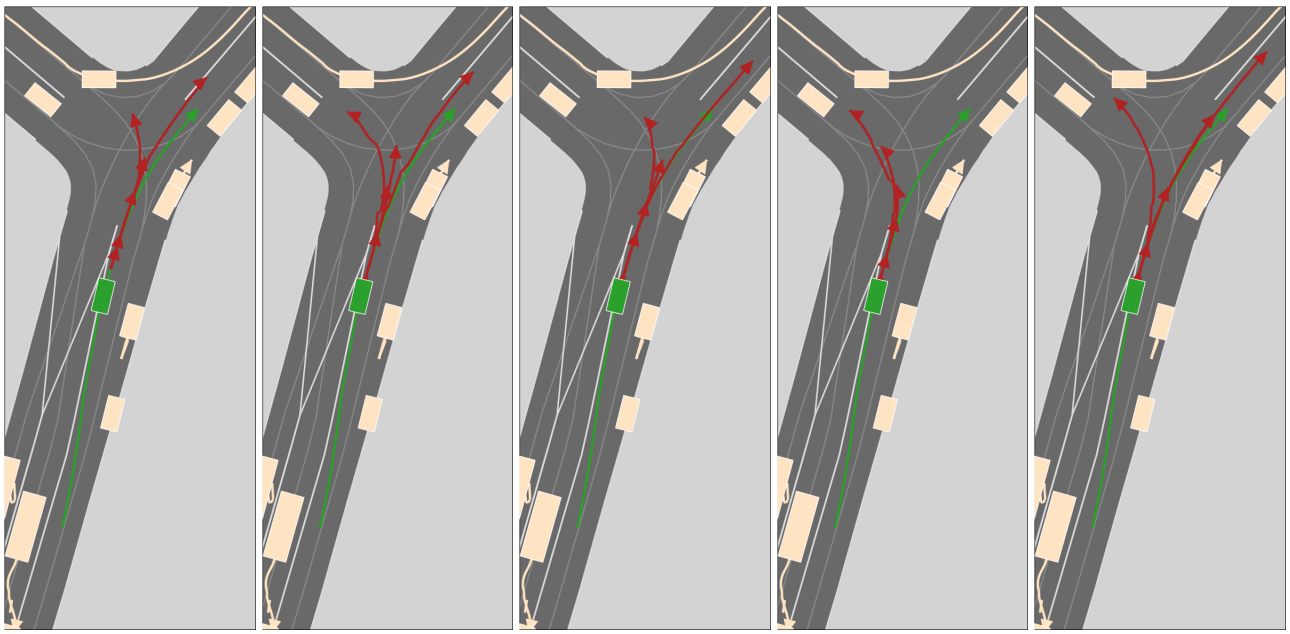
(a) Encoder-decoder (b) DONUT Overp. ✗ Ref. ✗ (c) DONUT Overp. ✓ Ref. ✗ (d) DONUT Overp. ✗ Ref. ✓ (e) DONUT Overp. ✓ Ref. ✓

Figure 11. **Additional qualitative results.**



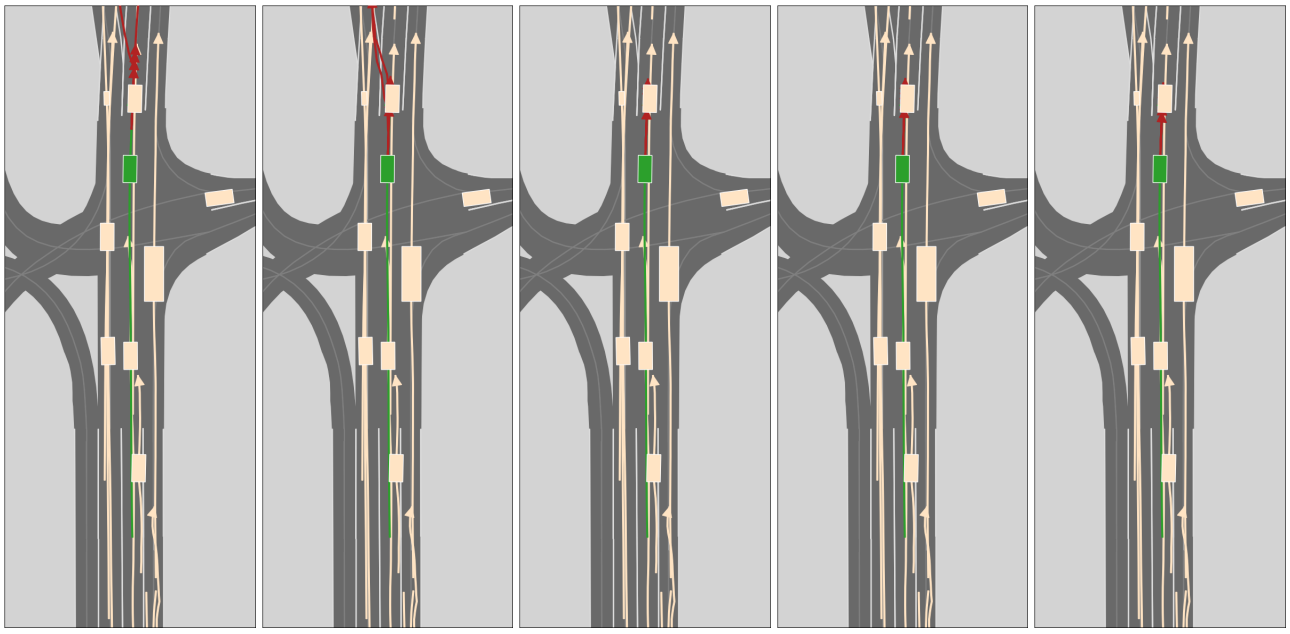
(a) Encoder-decoder (b) DONUT Overp. ✗ Ref. ✗ (c) DONUT Overp. ✓ Ref. ✗ (d) DONUT Overp. ✗ Ref. ✓ (e) DONUT Overp. ✓ Ref. ✓

Figure 12. **Additional qualitative results.**



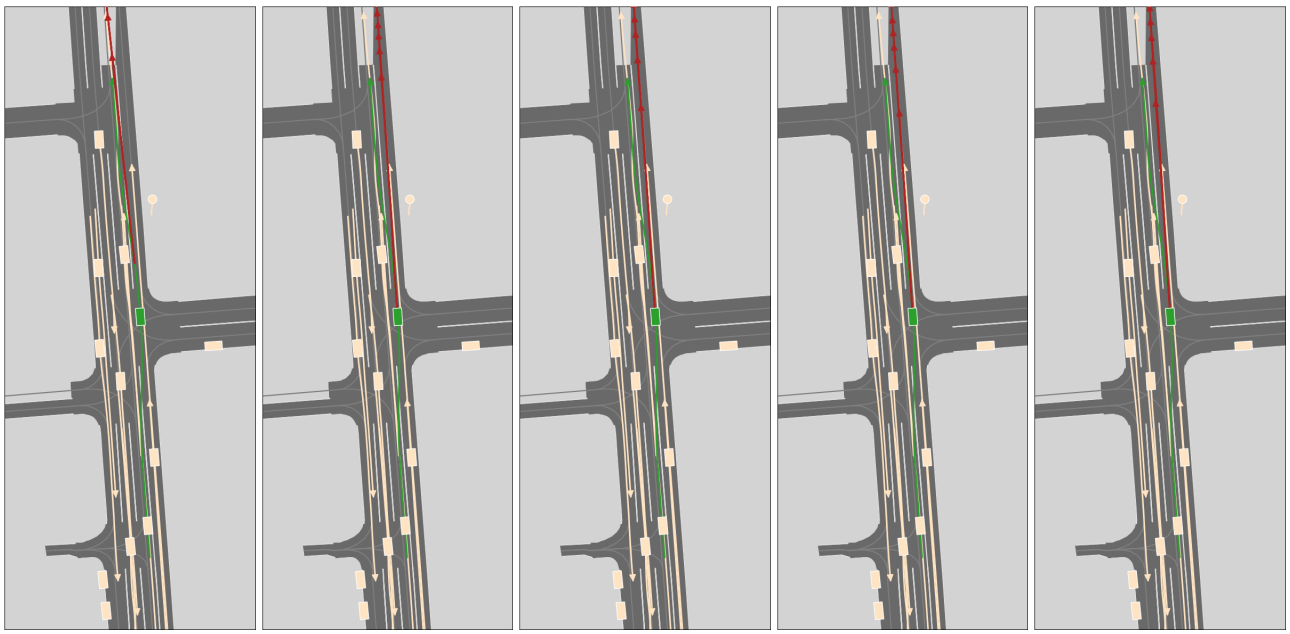
(a) Encoder-decoder (b) DONUT Overp. ✗ Ref. ✗ (c) DONUT Overp. ✓ Ref. ✗ (d) DONUT Overp. ✗ Ref. ✓ (e) DONUT Overp. ✓ Ref. ✓

Figure 13. **Additional qualitative results.**



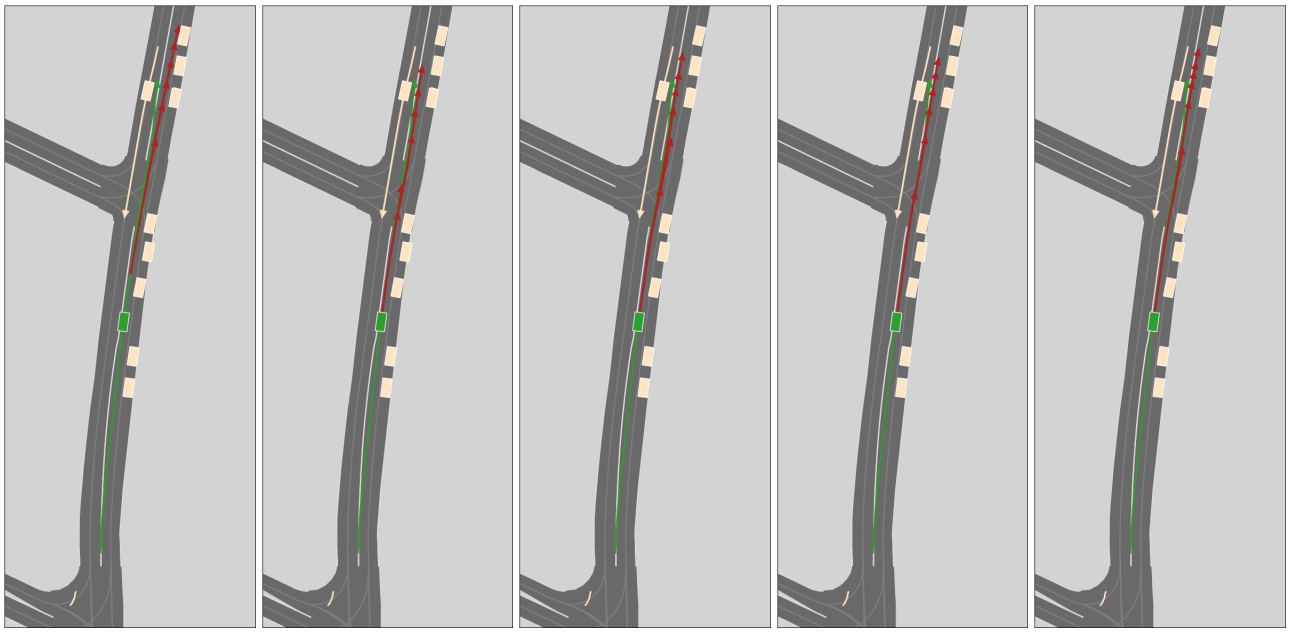
(a) Encoder-decoder (b) DONUT Overp. ✗ Ref. ✗ (c) DONUT Overp. ✓ Ref. ✗ (d) DONUT Overp. ✗ Ref. ✓ (e) DONUT Overp. ✓ Ref. ✓

Figure 14. **Additional qualitative results.**



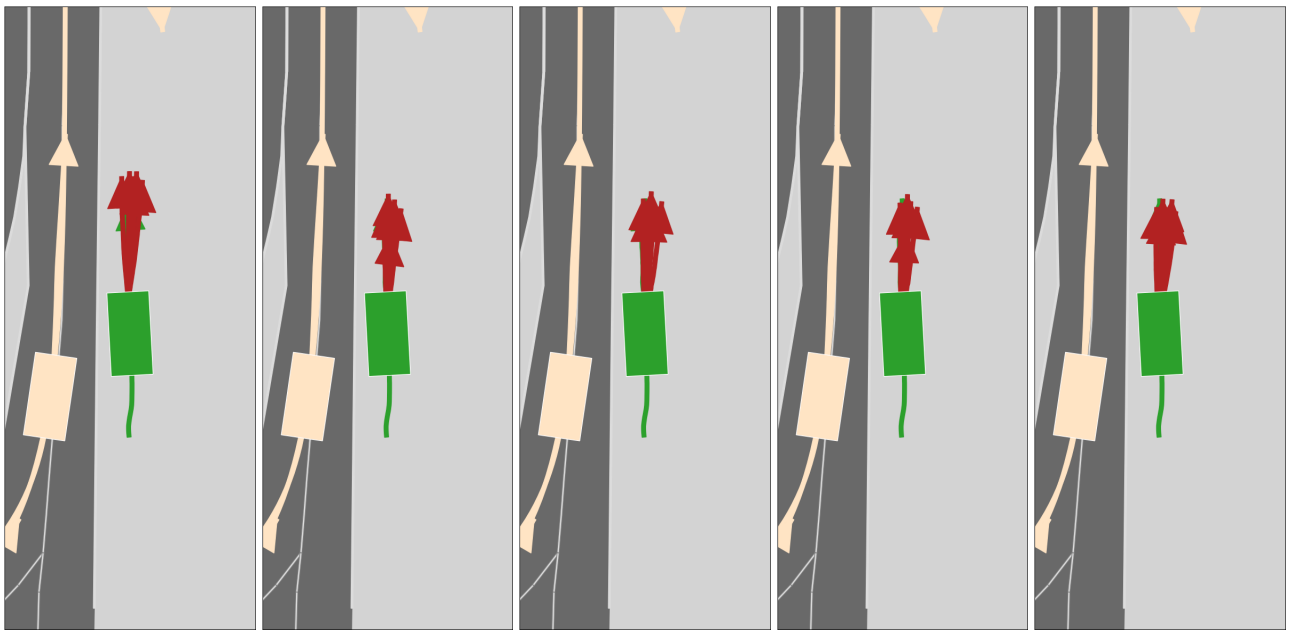
(a) Encoder-decoder (b) DONUT Overp. ✗ Ref. ✗ (c) DONUT Overp. ✓ Ref. ✗ (d) DONUT Overp. ✗ Ref. ✓ (e) DONUT Overp. ✓ Ref. ✓

Figure 15. **Additional qualitative results.**



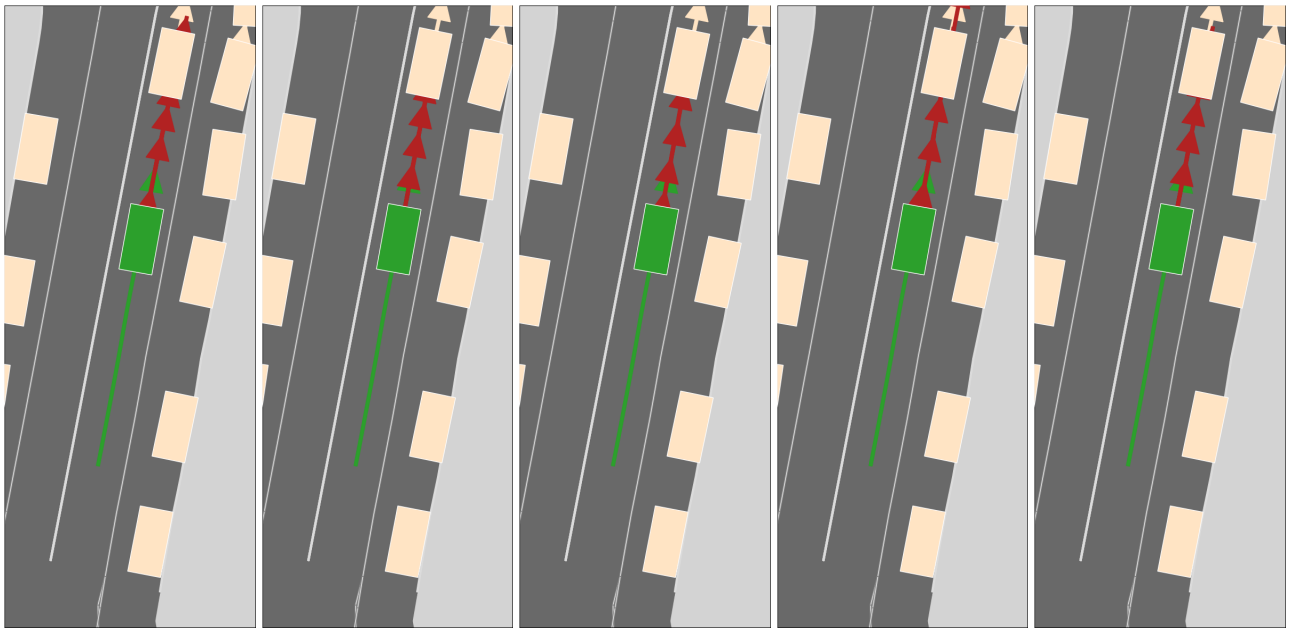
(a) Encoder-decoder (b) DONUT Overp. ✗ Ref. ✗ (c) DONUT Overp. ✓ Ref. ✗ (d) DONUT Overp. ✗ Ref. ✓ (e) DONUT Overp. ✓ Ref. ✓

Figure 16. **Additional qualitative results.**



(a) Encoder-decoder (b) DONUT Overp. ✗ Ref. ✗ (c) DONUT Overp. ✓ Ref. ✗ (d) DONUT Overp. ✗ Ref. ✓ (e) DONUT Overp. ✓ Ref. ✓

Figure 17. **Additional qualitative results.**



(a) Encoder-decoder (b) DONUT Overp. ✗ Ref. ✗ (c) DONUT Overp. ✓ Ref. ✗ (d) DONUT Overp. ✗ Ref. ✓ (e) DONUT Overp. ✓ Ref. ✓

Figure 18. **Additional qualitative results.**