Supplementary Material

1. The Detailed Network Architecture

Fig. 2 presents the detailed network architecture of our R-Pred. The entire network structure consists of ITPNet, TQSA module, PIA module and *prediction head*. TQSA and PIA take the output of ITPNet as an input and perform scene encoding and interaction encoding for trajectory refinement.

2. Additional Ablation Study

Input Formats of TRNet. TRNet is applied to the proposal features for refinement. The trajectory proposals are used only to extract the local scene features and conduct the distance-wise proposal grouping. To verify the advantage of our strategy, we compare our method with the baseline that re-encodes the trajectory proposals through MLP and applies the TRNet to the resulting features. Table 1 shows that by using the proposal features for refinement, our strategy outperforms the baseline by 1.87% and 4.4% in $minFDE_6$ and MR_6. This confirms the benefit of using the proposal features for the refinement network.

3. Scalability and Efficiency Analysis

Inference Speed of R-Pred. R-Pred is designed to predict the trajectories of all agents in a single forward pass by leveraging a multi-query decoding mechanism of Transformer. This approach ensures that scene features and proposal features are masked for each proposal using the tube-query pooling and proposal grouping algorithms. All proposals are fed into the Transformer as queries, producing the decoded features in parallel. To validate the effectiveness of our mechanism, we evaluated the impact of the number of agents on the inference speed of R-Pred using the Argoverse dataset. As illustrated in Fig. 1, inference for a single agent takes 201 ms, while predicting trajectories for 100 agents increases the inference speed by merely 44 ms. This demonstrates the scalability and effectiveness of our proposed method.

4. Additional Qualitative Examples

We visualize additional qualitative examples obtained in diverse interaction scenes. The examples were selected from *Argoverse validation set*. We compare the trajectory samples produced by ITPNet and TRNet to demonstrate the effectiveness of our refinement framework. We present the figures in two columns, where the left figures provide initial trajectory proposals, shown in blue, and the right figures provide the corresponding refined trajectories from TRNet, shown in red. The ground truth is shown as a green line and the trajectories of other agents are shown as black.

Speed Control Scenarios. In Fig. 3, we consider the scenarios where the target agents slow down or accelerate while interacting with other neighboring agent. In these examples, TRNet produces improved predictions by maintaining an appropriate distance from other agents.

Overtaking Scenarios. Fig. 4 shows three cases in which the target agents change lanes to overtake another agents. In all cases, TRNet produces trajectory predictions that are closer to the ground truth than ITPNet.

Intersection Scenarios. Fig. 5 shows the scenarios where the target agents interact with the nearby agents at intersections. Even if the initial trajectory proposals for two neighboring agents conflict, the trajectories in TRNet will not conflict after refinement.

Multi-modal Trajectory Behavior. Fig. 6 shows the multi-modal trajectory samples generated by ITPNet and TR-

Proposal feature	*Trajectory re-encoding*	$minADE_6$	$minFDE_6$	MR_6%
	✓	0.666	0.963	9.09
✓		**0.657**	**0.945**	**8.69**

Table 1. Comparison between our strategy and the baseline evaluated on the *Argoverse validation set*.

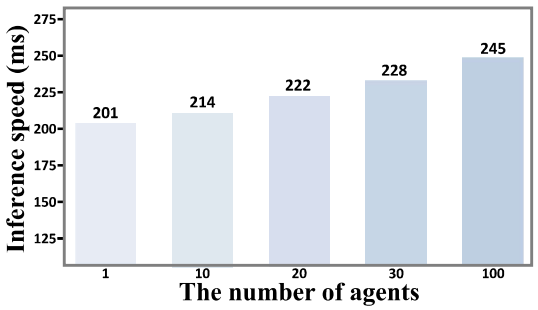

Figure 1. Inference speed based on the number of agents in the R-Pred on the *Argoverse validation set*.

Net. Some of ITPNet's trajectories do not seem plausible because they fall outside of road boundaries. In contrast, TRNet predicts the trajectories that better fit the scene structures and do not compromise the diversity of trajectory modes.

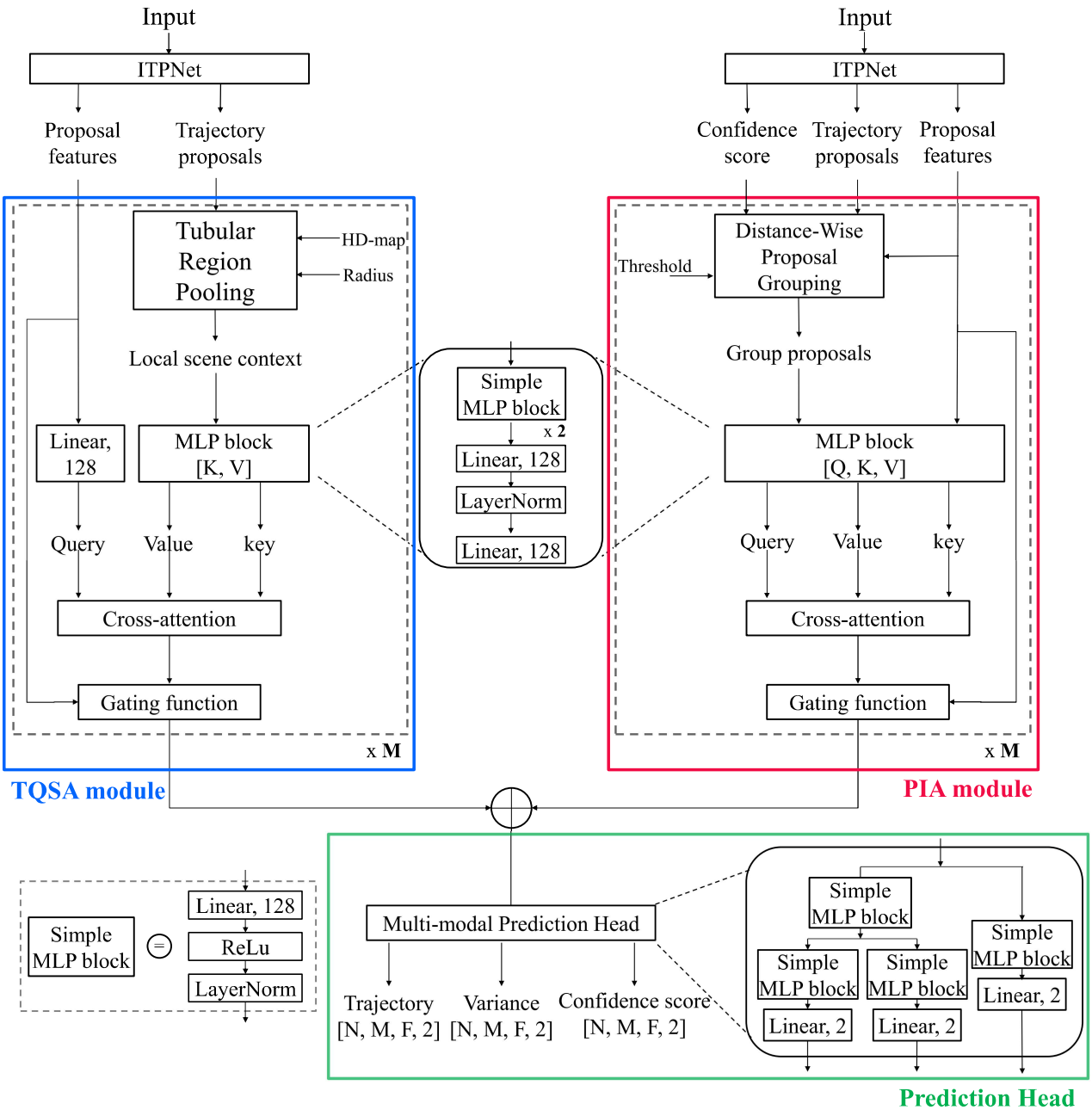

Figure 2. **Detailed architecture of R-Pred model.**

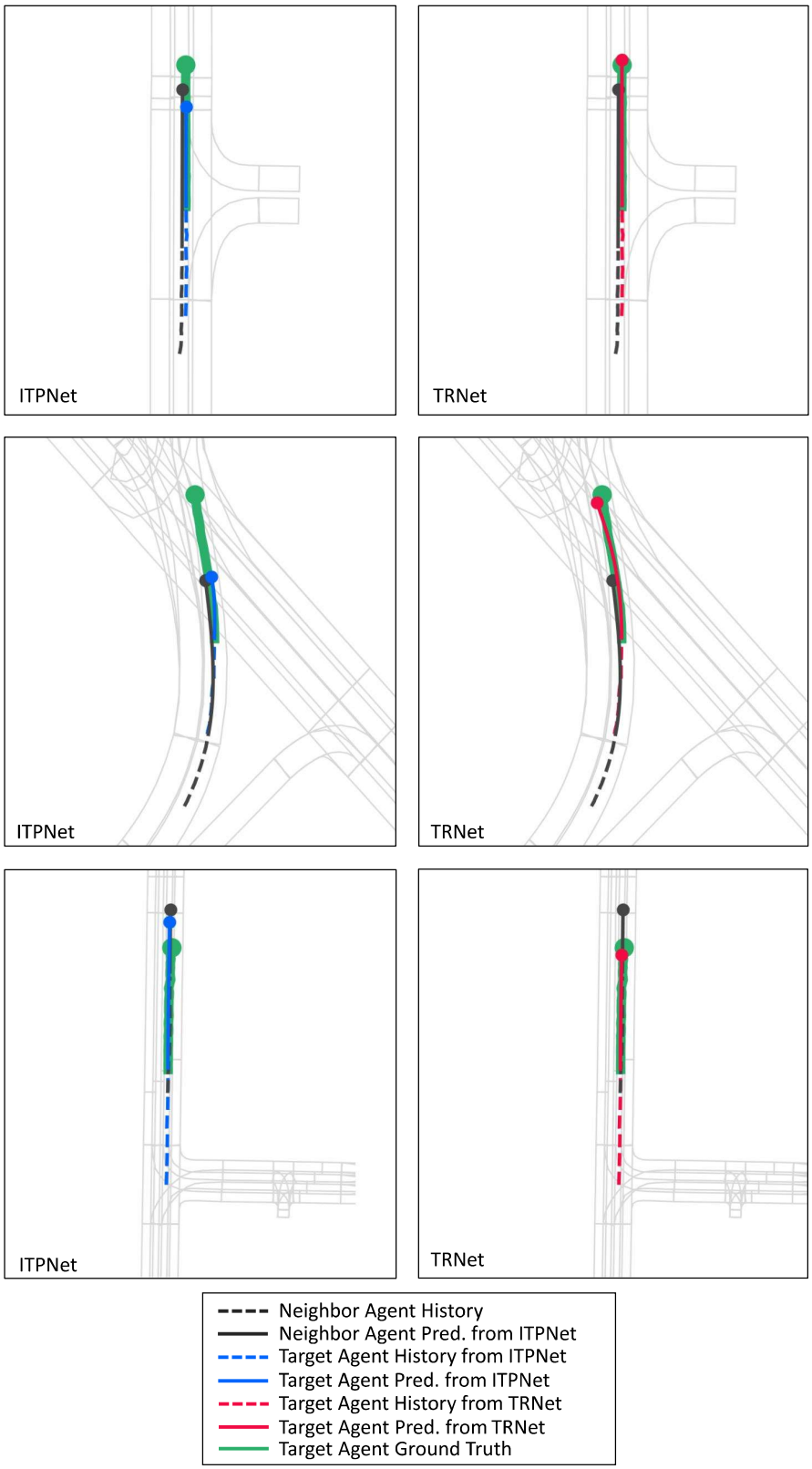

Figure 3. **Visualization of trajectories for several speed control scenarios.** In these scenarios, the target agents slow down or accelerate while interacting with other agents. The proposed refinement framework generates the predictions improved over the initial proposals from ITPNet.

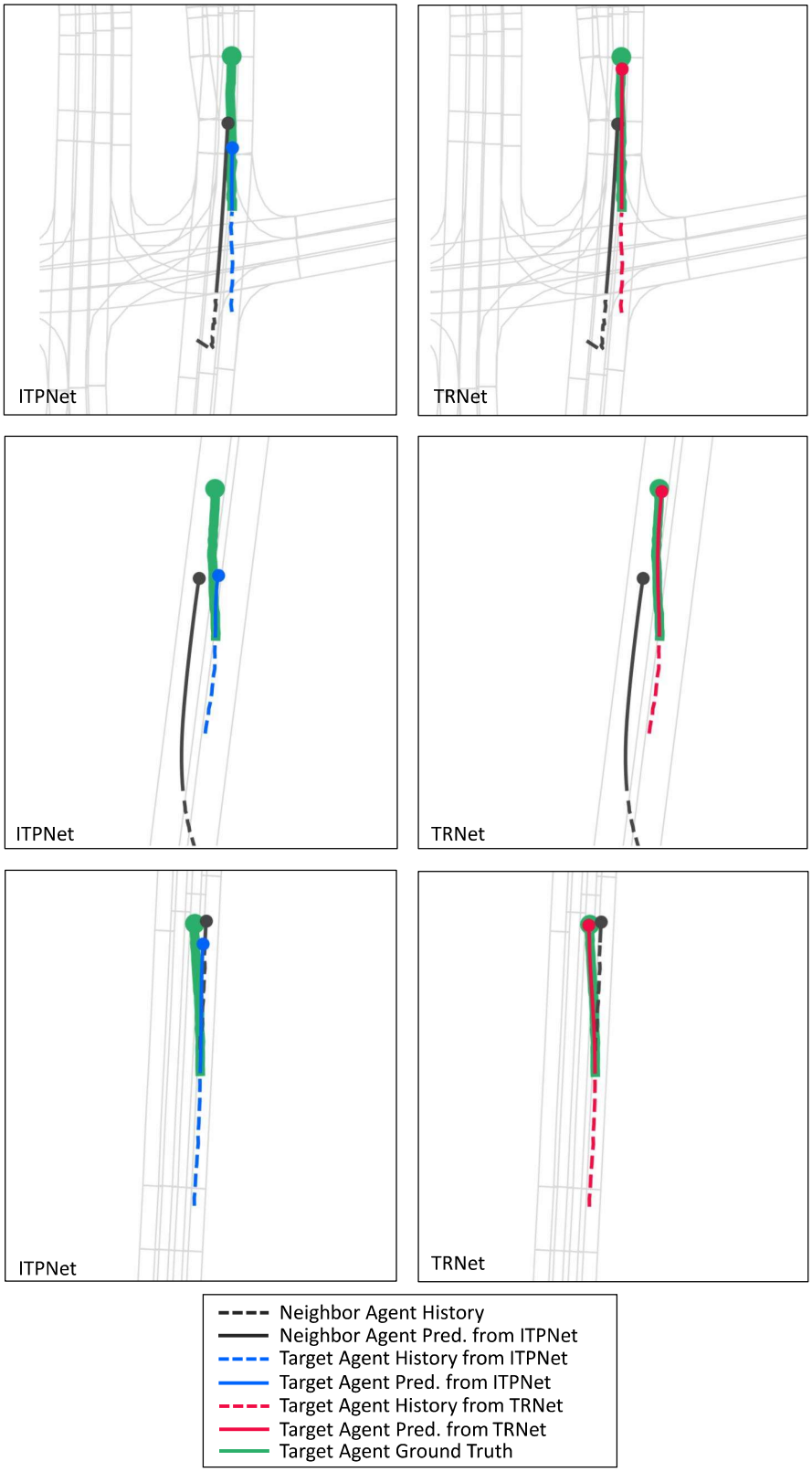

Figure 4. **Visualization of trajectories for several overtaking scenarios.** In these scenarios, the target agents change lanes to overtake other agents. Considering proposal-level interactions between the agents, TRNet produces trajectory predictions that are closer to the ground truth than ITPNet. Note that conflicts between the initial proposals from two neighboring agents are resolved by the proposed refinement framework.

Figure 5. **Visualization of trajectories for several intersection scenarios.** The target agents interact with the other agents at interactions. For all cases considered, TRNet produces trajectories that do not collide with other trajectories.

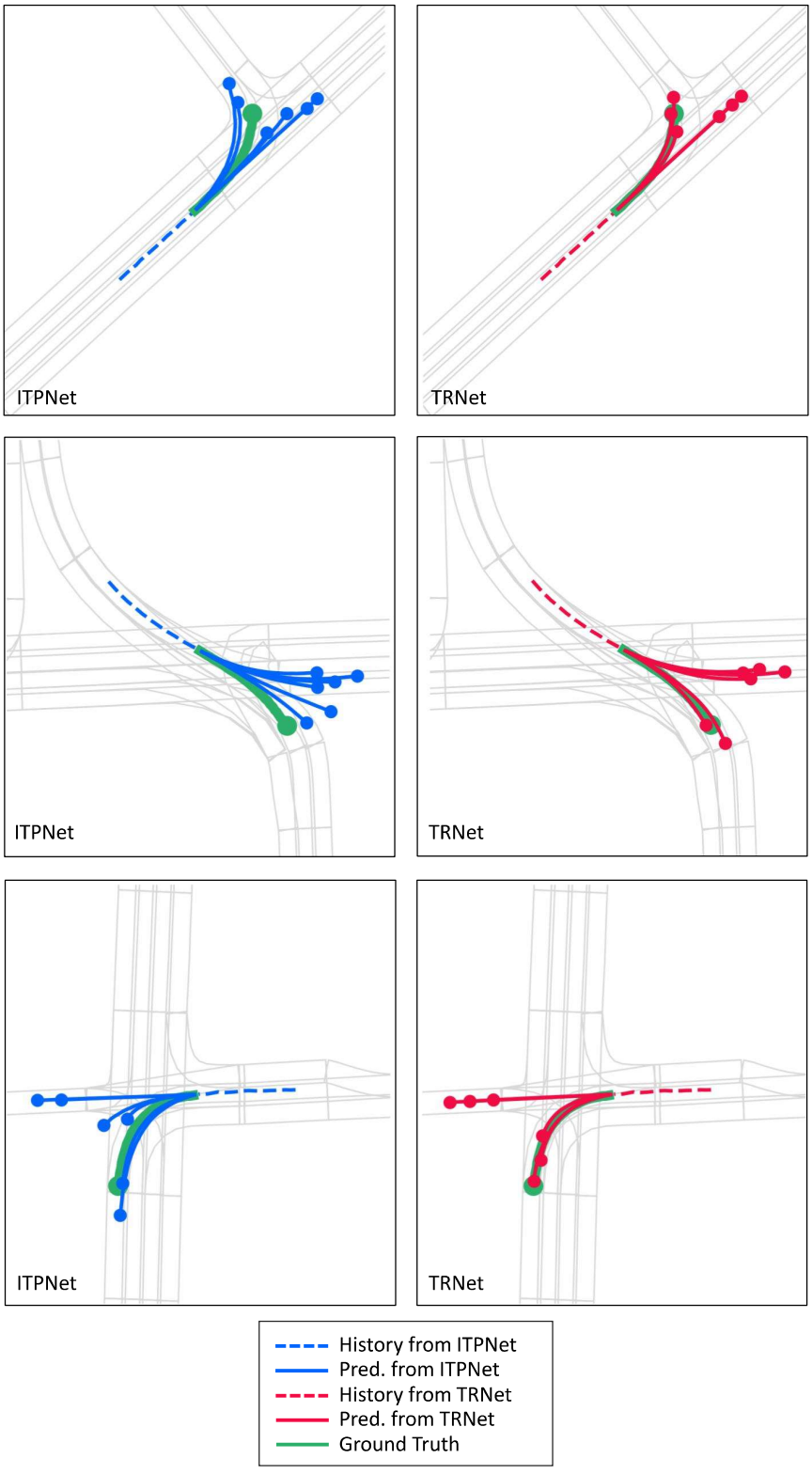

Figure 6. **Visualization of multi-modal trajectories obtained by ITPNet and TRNet.** The multi-modal trajectories predicted by TRNet mostly conform to the scene structures, whereas some trajectories generated by ITPNet are not physically plausible.