Appendix

A. Implementation Details

A.1. Predict Marginal Heatmap

The public source code of M2I [2] is used for predicting marginal heatmap. Specifically, we train three models for vehicle, pedestrian and cyclist separately. To encode scene context, we leverage the context encoder with both vectorized and rasterized representations. Please refer to M2I [2] and its source code for more details.

A.2. Data Preprocessing

Filter Interactive Pairs. Apart from the labeled interactive cases in the training set of WOMD, we filter more interactive pairs with the closest spatial distance [2] d_m in the future:

$$d_m = \min_{t_1=1}^{T} \min_{t_2=1}^{T} ||\boldsymbol{y}_1^{t_1} - \boldsymbol{y}_2^{t_2}||_2, \tag{1}$$

where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are the future trajectories of two agents with T steps. For each training case, we calculate d_m for any pair of agents and iteratively select pair with the smallest d_m over the left agents.

Prune Map with Marginal Heatmap. Given the marginal heatmap of each predicted agent, we prune the map limited by the area of top S intentions selected from heatmap using half circle and half ellipse. Concretely, we normalize intentions to the corresponding polyline coordinate of target agent, and calculate the distance between each intention and origin. d_f denotes the maximum distance over the intentions at the front of agent while d_r is the maximum distance over the intentions at the rear of agent. If $d_f > d_r$, the half circle with radius $r = d_f + 30$m is used at the front while the half ellipse with semi-major axis $a = r$ and semi-minor axis $b = d_r + 20$m is used at the rear. If $d_f \leq d_r$, the half circle with radius $r = d_r + 30$m is used at the rear while the half ellipse with semi-major axis $a = r$ and semi-minor axis $b = d_f + 20$m is used at the front. For the interactive cases, all road points within the local region of any target agent are reserved for trajectory prediction.

A.3. More Architecture Details

We train a single model for predicting all types of interactive pairs (any pairwise combination over vehicle, pedestrian and cyclist). Motivated by MTR [1], we use a three-layer MLP with dimension 256 to encode agent polylines,, and use a five-layer MLP with dimension 64 to encode road polylines. Both two types of polylines are further projected to dimension 256 with another linear layer separately. For the multi-modal decoder in HFIF (High-level Future Intentions Fusion), we use 1D convolution for goal regression and a three-layer MLP with dimension 256 for trajectory completion. For the multi-modal decoder in LFBF (Low-level Future Behaviors Fusion), we adopt a three-layer MLP with dimension 512 for trajectory prediction, and the weights are not shared across different layers.

A.4. Inference Latency

For the default setting of BiFF, the average inference latency is about $56ms$ for any case from Waymo interactive validation set. We measure the inference latency using a RTX 3090 GPU with standard Pytorch code.

B. Qualitative Results

The visualization results of our proposed method under complex interactive scenarios on the interactive validation set of WOMD are presented in Figure 1. The different interactive scenarios are shown in separate rows for clarity. In the first row, we demonstrate the effectiveness of our model in handling various types of agents by showcasing the interactions between vehicles and pedestrians. The middle row presents yielding scenarios among vehicles in complex intersections. Finally, the last row presents interactive merging scenarios where two agents are competing for the right-of-way at high speeds. These results illustrate the ability of our model to accurately predict long-term interactive scenarios in diverse and challenging scenarios. We also present more qualitative results for conflict resolution in Figure 2. Finally, we demonstrate the failure cases in Figure 3. The failures are mostly related to the misunderstanding of the agent's intention, for example, in the first figure, the blue agent is predicted to turn left in one of the modalities, then the model generates a more conservative behavior for red vehicle to yield the blue one.

C. Notations

To illustrate notations in the paper, Tab. 5 is provided.

D. Limitation and Future Work

We have identified several limitations of our proposed BiFF approach and outline potential avenues for future research. First, while BiFF demonstrates promising results, there remains a performance gap between our approach and SOTA in terms of mean average precision (mAP). We hypothesize that this discrepancy may be attributed to the inconsistency between the marginal heatmap and BiFF, as they are trained separately. To address this, one solution is to train a model that predicts the score of each agent separately, supervised with soft labels, and then obtain the joint score by multiplying the scores of all target agents in each modality. Second, limited by computing resources, the current version of BiFF is small without sufficient training data. We anticipate that increasing the amount of training data and decoder layers will continuously enhance the performance of BiFF. Third, the proposed approach can be

extended to handle more than two interactive targets with specifically designed techniques like sparse attention to reduce the computation of matrix multiplication.

E. Broader Impact

Regarding addressing real-world challenges for autonomous driving beyond the SOTA performance, we list our contributions to advance this area. First, from motivation, the joint trajectory prediction problem addressed by our BiFF is necessary for safe and comfortable driving, since a comprehensive understanding of the future trajectory distribution of all agents is more informative than marginal trajectory prediction. Second, thanks to polyline-based coordinates, our BiFF is memory-efficient without sacrificing accuracy, which is important for practical deployment and real-time inference. Third, apart from prediction task, the simulation and planning in self-driving will benefit from Bi-level Future Fusion by considering prediction as future for agents. Since the predicted trajectories are more scene-consistent with less unrealistic conflicting, BiFF is able to generate naturalistic driving behaviors for simulation and reduce collisions for planning.

References

[1] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *arXiv preprint arXiv:2209.13508*, 2022.

[2] Qiao Sun, Xin Huang, Junru Gu, Brian C Williams, and Hang Zhao. M2i: From factored marginal trajectory prediction to interactive prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6543–6552, 2022.

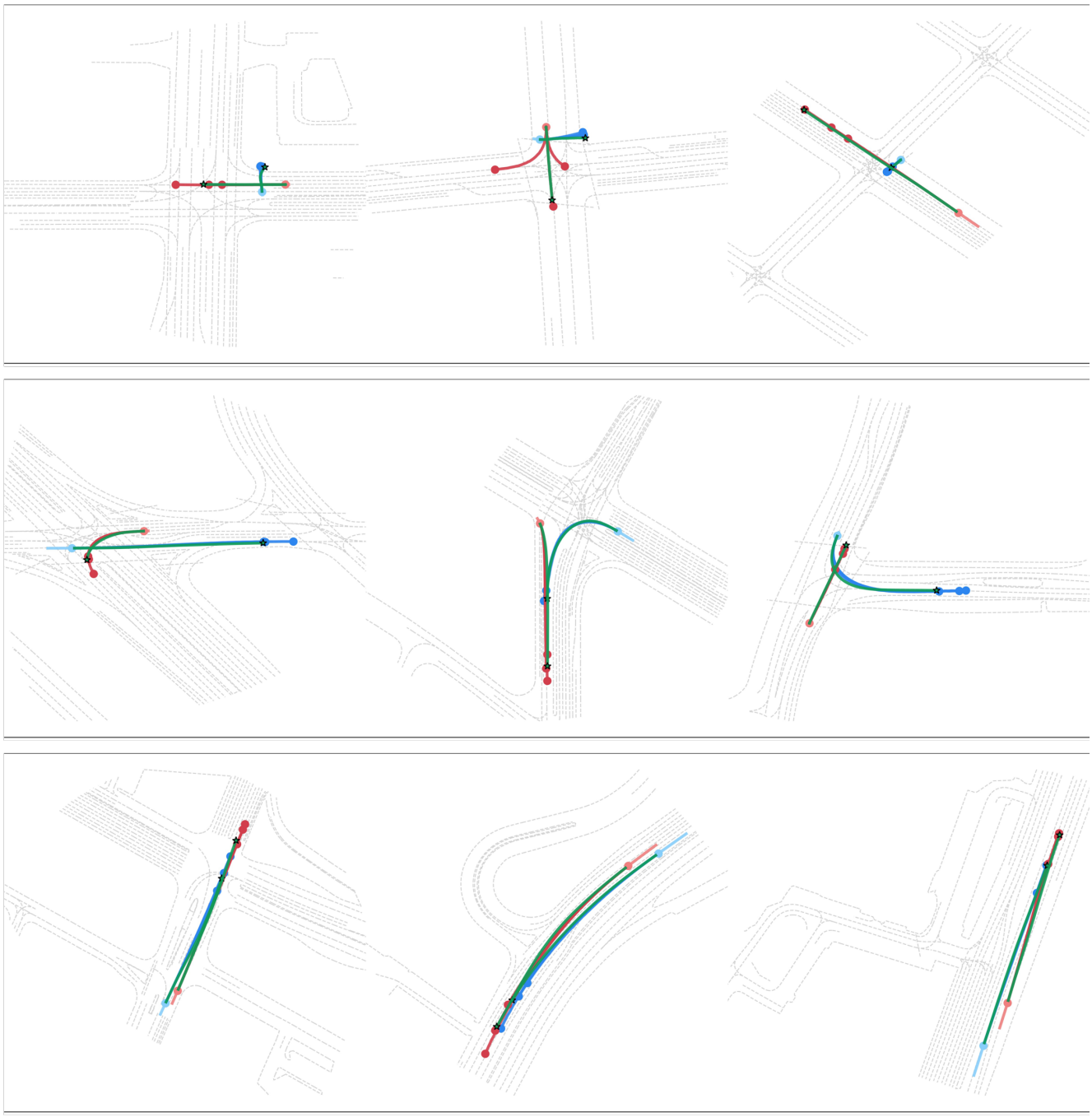

Figure 1. Qualitative results under diverse scenarios on the WOMD interactive validation set. HD map information is shown in light grey. For clarity, we choose K=3 pairs of predicted scene-compliant trajectories shown in red and blue while the corresponding history track is shown in a light color. Ground-truth future trajectories are illustrated in green on the top.

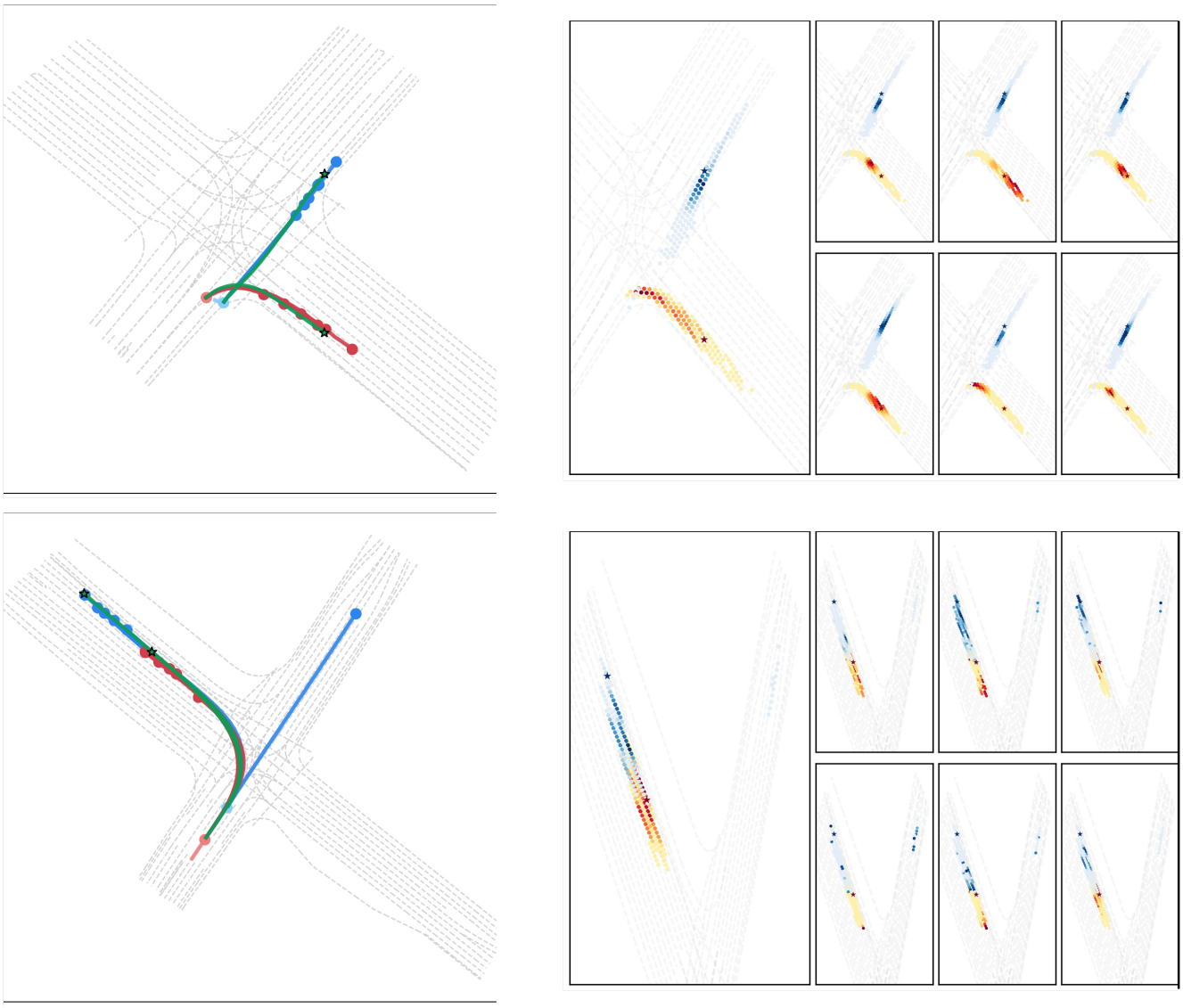

Figure 2. More qualitative results for conflict resolution. On the left column, we present the map information with predicted trajectories (K=6). On the right, we show the same scenario with heatmap representations. **Left**: Marginal heatmaps. **Right**: Scene-compliant assignment scores from different headers of motion decoder. Brightness in red and blue signifies the score of two interacting agents, with the asterisk denoting the ground truth goal points.

Figure 3. Qualitative analysis for failure cases on the WOMD interactive validation set.

Table 5. Lookup table for notations in the paper.

A	number of predicted interactive agents
S	number of static intentions (conditional anchors)
D	number of hidden feature dimension
K	number of predicted scene modalities
T	number of predicted future steps
L	number of nearest road polylines
N_E	number of stack layers of transformer encoder with relative positional encoding
N_L	number of stack layers of LFBF and multi-modal decoder
N_H	number of stack layers of HFIF
P_{ij}	relative positional encoding from polyline j to i
h_i^e	features of polyline i
d	the dimension of polyline feature
α	scaled dot-product attention
N_i	the set of polyline i's neighbors
$\gamma_{k,s}^a$	assignment scores for a-th agent in k-the modality
L_G	Goal regression
L_T	Trajectory regression
$\mathbf{y}^{1:T}$	Predicted future trajectory
N_{coll}	number of collision when calculating Cross Collision Rate