BiFF: Bi-level Future Fusion with Polyline-based Coordinate
for Interactive Trajectory Prediction

Yiyao Zhu   Di Luan   Shaojie Shen
The Hong Kong University of Science and Technology
{yzhucp, dluan}@connect.ust.hk, eeshaojie@ust.hk

Abstract

Predicting future trajectories of surrounding agents is essential for safety-critical autonomous driving. Most existing work focuses on predicting marginal trajectories for each agent independently. However, it has rarely been explored in predicting joint trajectories for interactive agents. In this work, we propose Bi-level Future Fusion (BiFF) to explicitly capture future interactions between interactive agents. Concretely, BiFF fuses the high-level future intentions followed by low-level future behaviors. Then the polyline-based coordinate is specifically designed for multi-agent prediction to ensure data efficiency, frame robustness, and prediction accuracy. Experiments show that BiFF achieves state-of-the-art performance on the interactive prediction benchmark of Waymo Open Motion Dataset.

1. Introduction

Motion prediction is crucial for intelligent driving systems as it plays a vital role in enabling autonomous vehicles to comprehend driving scenes and make safe plans in interactive environments. Predicting the future behaviors of dynamic agents is challenging due to the inherent multi-modal behavior of traffic participants and complex scene environments. Learning-based approaches [1, 3, 4, 7, 8, 12, 50, 9] have recently made notable strides in this area. By utilizing large-scale real-world driving datasets [2, 11], learning-based frameworks can effectively model complex interactions among agents and considerably enhance the accuracy of motion prediction.

Most existing studies on trajectory prediction [4, 12, 17, 22, 25] tend to focus on generating marginal prediction samples of future trajectories for each agent, without considering their future interactions. Thus, such marginal prediction models may generate trajectories with a high overlap rate [41], then the unreliable results are provided for the downstream planning module. To overcome this limitation, recent works [14, 16, 31] propose joint motion pre-
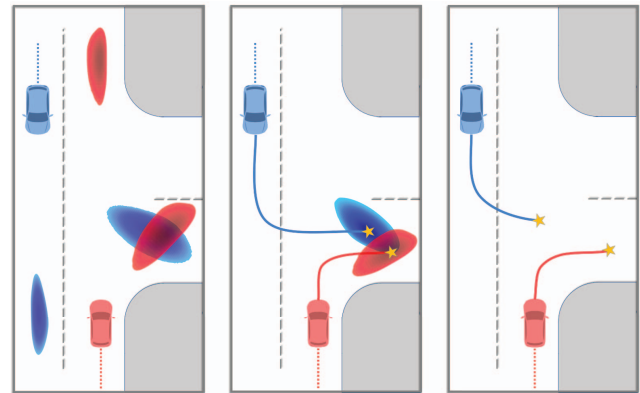

Figure 1. A motivating example of BiFF. **Left**: Marginal heatmaps of two interactive agents conflict with each other. **Middle**: For each scene modality, assignment scores generated by the prediction header become decoupled by fusing high-level future intentions across agents. **Right**: For each scene modality, the predicted trajectories are more scene-consistent by fusing low-level future behaviors across agents.

diction models that generate scene-compliant trajectories. However, previous models focus on capturing interaction across agents in the tracking history, explicitly modeling future interaction remains an open problem. The preliminary research regarding future interaction is conducted in the recent two works. M2I [41] classifies interactive agents as influencer and reactor. Conditioning on the influencer, the reactor benefit from future information, but the influencer is not optimized. MTR [38] predicts future trajectories for all the surrounding agents, but the future information is not reliable enough since trajectories are directly regressed using polyline features without any iterative refinement. To model the future interactions, we propose Bi-level Future Fusion (BiFF), and the motivating example is presented in Fig. 1, in which the unrealistic conflicting prediction between the marginal heatmaps is mitigated by the proposed High-level Future Intentions Fusion (HFIF) and Low-level Future Behaviors Fusion (LFBF).

Moreover, regarding these models [14, 16, 31], all target

agents are normalized to scene-centric coordinate, which sacrifices prediction accuracy and model generalization, therefore, requires heavy labor in data augmentation. One possible solution is to simply extract local features for agent-centric representation [52], but it is not memory-efficient with redundant context encoding between different target agents. To overcome these shortcomings, polyline-based coordinate is specifically designed to make all predictions invariant to the global reference frame, and the feature fusion between different coordinates is performed with relative positional encoding.

To summarize, our contributions remain as follows: (1) We propose a novel Bi-level Future Fusion (BiFF) model that incorporates a High-level Future Intention Fusion (HFIF) mechanism to generate scene-consistent goals and a Low-level Future Behavior Fusion (LFBF) mechanism to predict scene-compliant trajectories. (2) For multi-agent prediction, we design polyline-based coordinates to provide agent-centric representations for all target agents without redundant context encoding, which is memory-saving, data-efficient, and robust to the variance of global reference frame. (3) Our approach achieves state-of-the-art performance on the interactive prediction benchmark of Waymo Open Motion Dataset (WOMD).

2. Related Work

Recently, the rise in autonomous driving, as well as the availability of the corresponding datasets and benchmarks [2, 11, 48], have prompted significant research interest in motion prediction. Due to the uncertainty in human intent, numerous approaches have been explored to model the intrinsic multi-modal distribution. Earlier works employ stochastic models such as generative adversarial networks (GANS) [18, 36, 51] or conditional variational autoencoders (CVAES) [3, 22, 34, 42] to draw samples approximating the output distribution. However, sample inefficiency at inference time prohibits them from being deployed in safety-critical driving scenarios. Alternatively, some works, including [4, 20, 28, 37, 39], utilize Gaussian Mixture Models (GMMs) to parameterize multi-modal predictions and generate compact distributions. More recent work conduct classification on a set of predefined anchor trajectories [4, 32, 40] or potential intents such as goal targets [35, 13, 17, 50], such models have shown remarkable success in popular motion forecasting benchmark.

2.1. Scene Context Representation

Motion prediction models typically take road map information and agent history states as input. To encode such scene context efficiently, earlier works [1, 7, 8, 10, 18, 32] rasterize them into multi-channel BEV images for processing with convolutional neural networks (CNNs). Such models predominantly operate at a scene level and sacrifice the

pose-invariance of each target. In addition, recent studies have demonstrated the efficiency and scalability of encoding scene context using vectorized data directly from high-definition (HD) maps, as evidenced by the success of [12, 25, 46, 15, 43]. All targets still share the global frame in the vectorized methods above. By utilizing relative positional encoding, the agent-centric schemes [52, 6, 21] highly related to us are proposed, as they learn context representations that are invariant to the global transformation. However, our proposed method differs in three significant ways. (1) The redundant map information in HiVT [52] is precluded. (2) The feature encoder in our polyline-based coordinate is more straightforward than [52, 6, 21], where graph nodes aggregates features with hand-crafted design. (3) We incorporate future interactions in the agent-coordinate frame, similar to scene context encoding.

2.2. Interactive Trajectory Prediction

Multi-agent motion prediction is a crucial area, and one of the most challenging problems in this field is predicting scene-compliant future trajectories. To address this problem, agent interactions have been modelled using various techniques, such as hand-crafted functions [19], graph neural networks [25, 30, 24, 23] and attention mechanisms [40, 12, 28, 14, 5]. Recent studies have focused on using Transformers [44] to capture the interactions among agents and map elements [31, 16, 37, 26, 47]. While these models excel in modeling interactions in the encoding part, they tend to independently predict the final trajectory candidates losing scene-compliance.

In order to generate scene-consistent outputs, ILVM [3] uses scene latent variables for interaction modeling and optimizes with ELBO. SceneTransformer [31] and AutoBots [16] decode joint motion predictions with learnable queries in the Transformer block. Thomas [14] directly designed a recombination model to recombine the agents' endpoints generating joint futures. M2I [41] identifies an influencer-reactor pair. To explicitly capture future interaction, the trajectory of reactor is predicted depending on influencer. JFP [27] utilizes a dynamic interactive graph model on top of the trajectory candidates to generate consistent future predictions. MTR [38] presents an auxiliary dense future prediction task to facilitate future interaction modelling and guide the model to generate scene-compliant trajectories.

3. Approach

3.1. Model Overview

The overall framework of BiFF is illustrated in Fig. 2. For the scene encoder, we represent the driving scene in a vectorized format for both agents and roads normalized to their own polyline-based coordinate separately, which achieves agent-centric representation for predicted agents
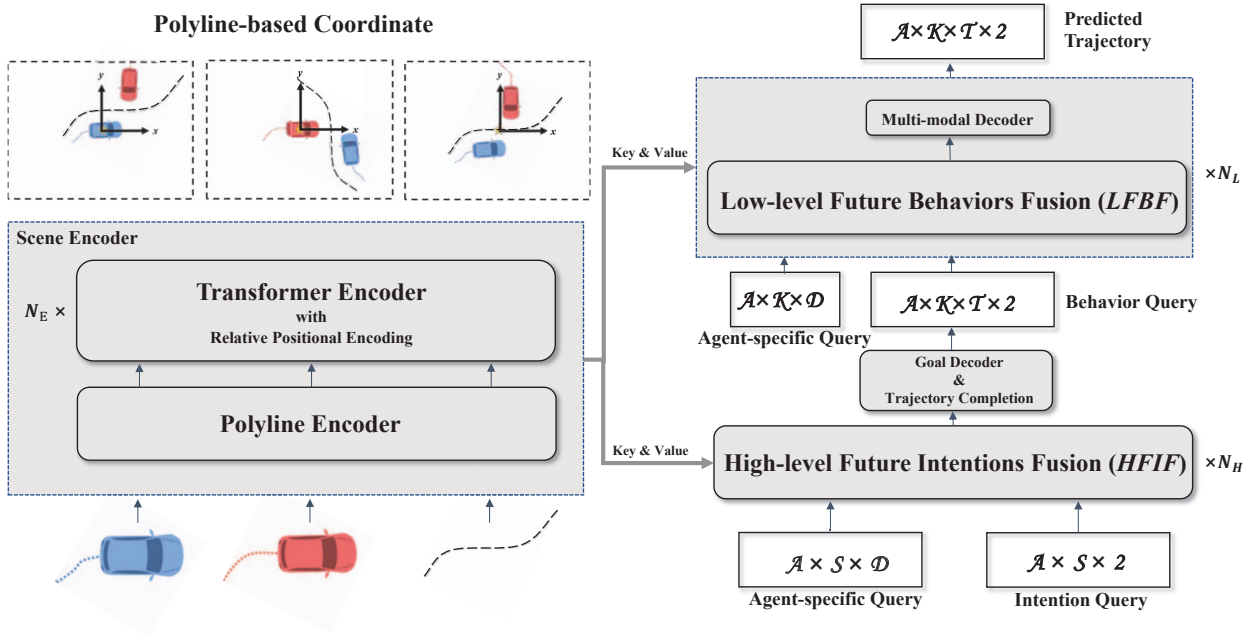
Figure 2. BiFF framework overview. Left indicates the scene context encoder, with agents (in red and blue) and road (in dashed curves) normalized to their own polyline-based coordinate separately. Right shows the motion decoder, with A predicted interactive agents, S static intentions (conditional anchors), D hidden feature dimension, K predicted scene modalities, and T predicted future steps. N_E, N_L and N_H are the number of stacked layers. Details can be referred to Model Overview (Sec.3.1).

without any redundant context encoding (Sec.3.2). Then agents and roads are encoded by a PointNet-like [33] poly-line encoder followed by transformer encoder with relative positional encoding. For the motion decoder, the High-level Future Intentions Fusion (HFIF) module fuses static future intentions across target agents followed by multi-modal goal decoder and trajectory completion (Sec.3.3). Given the predicted trajectory from HFIF, we take the Low-level Future Behaviors Fusion (LFBF) module to fuse dynamic future behaviors of interactive agents in each scene modality (Sec.3.4). Finally, we present the details of training loss and model inference (Sec.3.5).

3.2. Scene Encoder with Polyline-based Coordinate

For the joint trajectory prediction, existing works apply scene-centric representation normalizing all context to a unique global frame [14, 16, 31]. We argue that polyline-based coordinate with relative positional encoding is more efficient for interactive prediction.

Represent context with polyline-based coordinate. We adopt vectorized method proposed in [12] to represent agents and roads as polylines. The agent-centric strategy is utilized with polyline-based coordinate. For the pose normalization, each agent is centered to their own pose at the current time step while each road is normalized to the cen-

ter of its own polyline. With the normalized features, for the polyline construction, each agent is composed of all history frames with state features (e.g. position, velocity, heading, type, etc) while each road consists of maximum 10 points sampled at intervals of 1 meter with point information (e.g., position, direction, type, etc.). Finally, we use PointNet-like [33] architecture with an MLP followed by max pooling per polyline to obtain the feature per element of agent or road.

Fuse polylines with relative positional encoding. We leverage transformer encoder stacked with N_E layers to fuse features across polylines. Considering that all polylines are self-normalized, to make each polyline aware of the poses of others, the relative positional encoding is utilized for performing pair-wise message passing. We take the directional feature fusion between polylines $\mathbf{p}_i$ and $\mathbf{p}_j$ (i.e., $j \to i$) as an example, where the pairwise relative positional encoding $\mathbf{p}_{ij}$ is obtained using MLP to the relative pose (e.g., $\Delta x_{ij}, \Delta y_{ij}, \cos(\Delta\theta_{ij}), \sin(\Delta\theta_{ij})$). Concretely, by normalizing the coordinate of polyline j to polyline i, $(\Delta x_{ij}, \Delta y_{ij})$ is the origin location and $\Delta\theta_{ij}$ is the rotation angle of polyline j in the coordinate of polyline i. Then we incorporate $\mathbf{p}_{ij}$ into the vector transformation. The polyline i is used to calculate query vector, and the transformed polyline j is considered as key and value vectors:

$$\mathbf{q}_i^e = \mathbf{W}_q^e \mathbf{h}_i^e,$$
$$\mathbf{k}_{ij}^e = \mathbf{W}_k^e \mathbf{h}_j^e + \mathbf{W}_{pos}^e \mathbf{p}_{ij}, \quad (1)$$
$$\mathbf{v}_{ij}^e = \mathbf{W}_v^e \mathbf{h}_j^e$$

where $\mathbf{h}_i^e$ and $\mathbf{h}_j^e$ are the features of polyline i and polyline j respectively, $\mathbf{W}_q^e, \mathbf{W}_k^e, \mathbf{W}_v^e, \mathbf{W}_{pos}^e \in \mathbb{R}^{d \times d}$ are learnable matrices for linear projection, and d is the dimension of polyline feature. The weight matrices are not shared across different encoder layers. The scaled dot-product attention is calculated by the given query, key and value:

$$\alpha_{ij}^e = \frac{\exp\left(\mathbf{q}_i^{e\top} \mathbf{k}_{ij}^e / \sqrt{d_k}\right)}{\sum_{j \in \mathcal{N}_i} \exp\left(\mathbf{q}_i^{e\top} \mathbf{k}_{ij}^e / \sqrt{d_k}\right)}, \quad (2)$$

$$\hat{\mathbf{h}}_i^e = \sum_{j \in \mathcal{N}_i} \alpha_{ij}^e \mathbf{v}_{ij}^e, \quad (3)$$

where $\mathcal{N}_i$ is the set of polyline i's neighbors with k closest polylines (whose polyline coordinates are closest to the query polyline i), d_k is the dimension of query or key. The updated feature $\hat{\mathbf{h}}_i^e$ output from the final encoder layer is considered as polyline feature for the motion decoder. For simplicity, we omit the commonly used Multi-Head Attention, Layer Normalization, Feed-Forward Network and Residual Connection of Transformer [44] in the equations.

3.3. High-level Future Intentions Fusion (HFIF)

Given the marginal target heatmap of each predicted agent [1], we select top S intentions with the highest scores as conditional anchors for High-level Future Intentions Fusion (HFIF). As shown in the left part of Fig. 3, the transformer decoder with static query intentions is utilized to localize each goal and aggregate intention-specific context. To fuse the high-level future intentions, the decoder is further extended with HFIF. Finally, K group goals are generated with multi-modal decoder followed by a simple trajectory completer. Note that in both HFIF (Sec.3.3) and LFBF (Sec.3.4), all features fusion between different coordinates are performed with relative positional encoding, and the learnable weights are not shared across different decoder layers. The HFIF is stacked with N_H layers, and the detailed architecture is illustrated in the following.

Intention queries and attention module. In the HFIF-based transformer decoder layer, to distinguish queries of different agents, we initialized them using the fused feature $\mathbf{h}^e$ of corresponding target agent obtained from the scene encoder (Sec.3.2) in the first decoder layer, but using the query content feature output from previous layer in the following decoder layers. Then the agent-specific queries are

[1]The details of model architecture and training for predicting marginal heatmap can be found in Appendix.

further augmented with the corresponding static query intentions in all decoder layers. Given S intentions normalized to the predicted agent's polyline coordinate, we apply MLP to encode each query intention $\mathbf{s} \in \mathbb{R}^2$ (x, y) as $\mathbf{e}^{qi} \in \mathbb{R}^d$ and add it to the query content. Then the self-attention layer ($\mathbf{I_{SA}}$) is utilized to propagate information among S queries inner each agent, and the updated intention query $\mathbf{h}^{isa} \in \mathbb{R}^d$ is obtained.

Next, we introduce cross-attention block ($\mathbf{I_{CA}}$) with relative positional encoding to aggregate intention-specific context. For example, any intention of any target agent i is used as the query vector, and any polyline j is considered as key and value vectors. Motivated by [38], to decouple the attention contribution from two types of features, we concatenate query content with intention embedding while concatenate key content with positional encoding:

$$\mathbf{q}_i^{ica} = \left[\mathbf{W}_q^{ica} \mathbf{h}_i^{isa}, \mathbf{W}_{qe}^{ica} \mathbf{e}_i^{qi}\right],$$
$$\mathbf{k}_{ij}^{ica} = \left[\mathbf{W}_k^{ica} \mathbf{h}_j^e, \mathbf{W}_{pos}^{ica} \mathbf{p}_{ij}\right], \quad (4)$$
$$\mathbf{v}_{ij}^{ica} = \mathbf{W}_v^{ica} \mathbf{h}_j^e,$$

where the intention feature $\mathbf{h}_i^{isa}$ is returned by self-attention module paired with the intention embedding $\mathbf{e}_i^{qi}$, the polyline feature $\mathbf{h}_j^e$ is output from scene encoder, and $\mathbf{p}_{ij}$ is the relative positional encoding (normalize polyline coordinate of polyline j to target agent i). All $[\cdot, \cdot]$ denotes concatenating two vectors $\in \mathbb{R}^d$ as one vector $\in \mathbb{R}^{2d}$ and all $\mathbf{W}$ (omit superscript and subscript) are learnable matrices from Eq. 4 to Eq. 7. Considering that information aggregation from agent or road is inherently distinguished in the attention mechanism, we adopt two groups of projection matrices for agent and road separately in both HFIF (Sec.3.3) and LFBF (Sec.3.4). By calculating cross attention using Eq. 2 and Eq. 3, for any intention of any target agent i, the updated query intention $\mathbf{h}_i^{ica} \in \mathbb{R}^d$ is obtained. The $\mathcal{N}_i$ in Eq. 2 is composed of features of all agent polylines and the closest L road polylines (whose centers are closest to the current query intention embedded as $\mathbf{e}_i^{qi}$) output from scene encoder.

Fuse high-level future intentions across agents. We introduce **HFIF** module to capture the interaction of future intentions across target agents. Fig. 4 shows the details of HFIF, where any intention of any target agent aggregates features of all static intentions from other targets using cross-attention. For example, when performing message passing from any intention of target agent j to any intention of target agent i, the fused intention features paired with their original embedding are incorporated into the attention mechanism with relative positional encoding $\mathbf{p}_{ij}$ (normalize polyline coordinate of target agent j to target agent i):
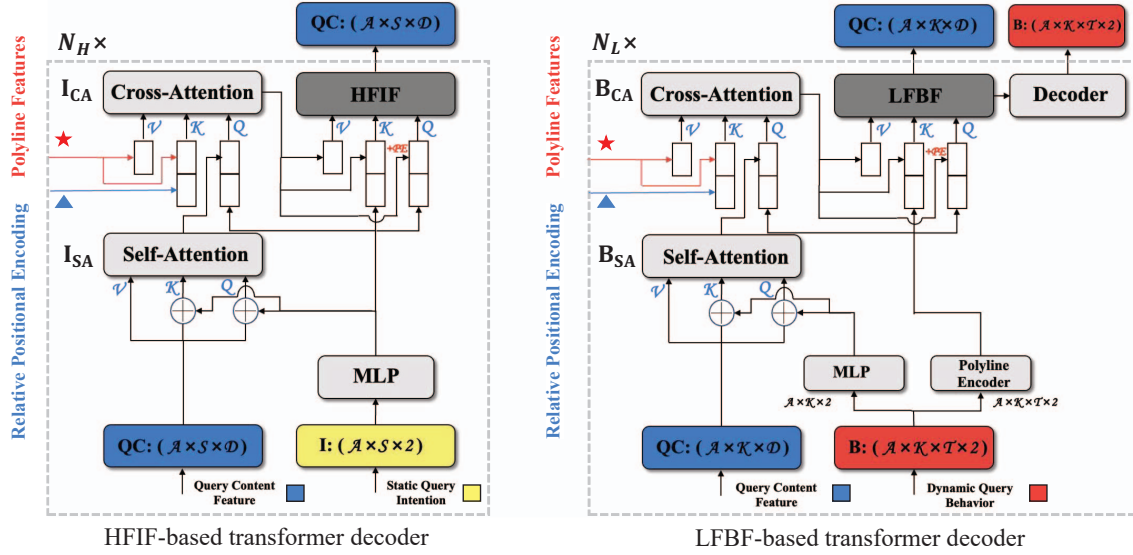
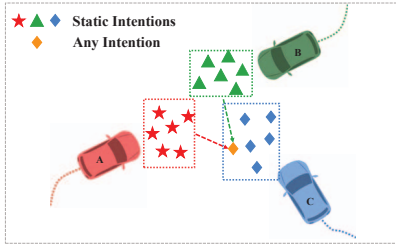Figure 3. Details of HFIF-based and LFBF-based decoder network structure.

Figure 4. HFIF: any intention aggregates features of all static intentions from other target agents.
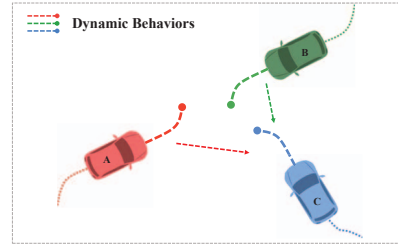

Figure 5. LFBF: any behavior aggregates features of all dynamic behaviors in the same scene modality from other target agents.

$$\mathbf{q}_i^{hif} = \left[\mathbf{W}_q^{hif}\mathbf{h}_i^{ica}, \mathbf{W}_{qe}^{hif}\mathbf{e}_i^{qi}\right],$$

$$\mathbf{k}_{ij}^{hif} = \left[\mathbf{W}_k^{hif}\mathbf{h}_j^{ica} + \mathbf{W}_{pos}^{hif}\mathbf{p}_{ij}, \mathbf{W}_{ke}^{hif}\mathbf{e}_j^{qi}\right], \quad (5)$$

$$\mathbf{v}_{ij}^{hif} = \mathbf{W}_v^{hif}\mathbf{h}_j^{ica},$$

where $\mathbf{h}_i^{ica}$ and $\mathbf{h}_j^{ica}$ output from cross attention module is the scene-based features of any intention of target agent i and agent j, while $\mathbf{e}_i^{qi}$, $\mathbf{e}_j^{qi}$ are the corresponding intention embedding, respectively. To make $\mathbf{e}_i^{qi}$ aware of the position of $\mathbf{e}_j^{qi}$, the location of any intention of agent j is normalized to the polyline coordinate of agent i before embedding. The computation in Eq. 2 and Eq. 3 is conducted for the high-level intention fusion (hif) with the neighbor $\mathcal{N}_i$ containing all intention features of other target agents. Finally, we obtain the updated feature $\mathbf{h}_i^{hif} \in \mathbb{R}^d$ for any intention of any target agent i, and the feature is used as the initialized query content for the next decoder layer.

Multi-modal decoder in HFIF. Given $\mathbf{h}_i^{hif}$ output from the final decoder layer in HFIF, we introduce the multi-modal decoder composed of goal regression followed by trajectory completion. To extract different interactive modes, we apply a multi-modal goal decoder containing $K = 6$ headers to the fused intention features from HFIF-based transformer decoder, and all headers are consist of 1D convolution with identical structure. Each header predicts scene-consistent goals for all target agents in each scene modality. For example, to regress goals of the a-th agent in the k-th modality, all fused intention features of the a-th agent are passed through the k-th header followed by a Softmax function to predict the assignment scores $\gamma_{k,s}^a$ (s is the s-th intention in all S conditional anchors). The corresponding regressing goal $\hat{g}_k^a$ can be computed by the weighted sum of the conditional anchors' coordinate p_s^a, i.e., $\hat{g}_k^a = \sum_{s=1}^{S}\gamma_{k,s}^a p_s^a$. Given the goals predicted for each scene modality, we concatenate each goal with the corre-

sponding agent feature from scene encoder (Sec.3.2), then pass the new embedding to an MLP to complete predicted trajectory for each target in each scene modality.

3.4. Low-level Future Behaviors Fusion (LFBF)

After HFIF, we obtain K groups of predicted trajectories considered as dynamic anchors for Low-level Future Behaviors Fusion (LFBF). Refer to the right part of Fig. 3, the transformer decoder with dynamic query behaviors (predicted trajectories) is utilized to localize each modality and aggregate behavior-specific context. We extend decoder with LFBF to fuse the low-level future behaviors in each scene modality. The LFBF-based transformer decoder is stacked with N_L layers to iteratively refine predicted trajectories and fuse future behaviors. The details are illustrated as follows.

Behavior queries and attention module. In the LFBF-based transformer decoder layer, the dynamic query behaviors are initialized by the trajectories generated from HFIF, and dynamically updated with the trajectories predicted from previous decoder layer, which is designed for iteratively trajectory refine [38]. Given any future trajectory $\mathbf{y}^{1:T} \in \mathbb{R}^{T \times 2}$ normalized to the corresponding polyline coordinate of any target agent, we encode the query behavior as $\mathbf{e}^{qb} \in \mathbb{R}^d$ using polyline encoder (a small PointNet-like [33] structure) and encode the endpoint $\mathbf{y}^T \in \mathbb{R}^2$ as $\mathbf{e}^{qeb} \in \mathbb{R}^d$ using MLP. The $\mathbf{e}^{qb}$ is applied to the scene context aggregation and future behaviors fusion while $\mathbf{e}^{qeb}$ is utilized to localize each modality. Regarding initializing query content in each decoder, we take the same process in HFIF, and the queries are added with dynamic $\mathbf{e}^{qeb}$ to distinguish each modality in all decoder layers. For the self-attention layer ($\mathbf{B_{SA}}$), we apply it to propagate information among K modalities inner each agent, and the updated behavior query $\mathbf{h}^{bsa} \in \mathbb{R}^d$ is obtained.

Next, we introduce cross-attention block ($\mathbf{B_{CA}}$) with relative positional encoding to aggregate behavior-specific scene context. For example, any future behavior of any target agent i is used as the query vector, and any polyline j is considered as key and value vectors:

$$
\begin{aligned}
\mathbf{q}_i^{bca} &= \left[\mathbf{W}_q^{bca} \mathbf{h}_i^{bsa}, \mathbf{W}_{qe}^{bca} \mathbf{e}_i^{qb} \right], \\
\mathbf{k}_{ij}^{bca} &= \left[\mathbf{W}_k^{bca} \mathbf{h}_j^e, \mathbf{W}_{pos}^{bca} \mathbf{p}_{ij} \right], \\
\mathbf{v}_{ij}^{bca} &= \mathbf{W}_v^{bca} \mathbf{h}_j^e,
\end{aligned}
\tag{6}
$$

where the behavior feature $\mathbf{h}_i^{bsa}$ is returned by self-attention module paired with the behavior embedding $\mathbf{e}_i^{qb}$, polyline feature $\mathbf{h}_j^e$ is output from scene encoder, and $\mathbf{p}_{ij}$ is the relative positional encoding (normalize coordinate of polyline j to target agent i). For any target agent i in any scene modality, we update the query behavior as $\mathbf{h}_i^{bca} \in \mathbb{R}^d$ by calculating cross attention with Eq. 2 and Eq. 3, where

$\mathcal{N}_i$ is composed of features of all agent polylines and the closest L road polylines (whose centers are closest to the center of the current query behavior embedded as $\mathbf{e}_i^{qb}$).

Fuse low-level future behaviors across agents. In each scene modality, we introduce **LFBF** module to capture the interaction of low-level future behaviors across target agents. Fig. 5 illustrates the details of LFBF, where any behavior of any target agent aggregates features of all dynamic behaviors in the same scene modality from other targets using cross-attention. For example, when performing message passing from future behavior of target agent j to target agent i in any scene modality, the fused future behavior features paired with their trajectory embedding are introduced to calculate query, key and value vectors with relative positional encoding $\mathbf{p}_{ij}$ (normalize polyline coordinate of target agent j to target agent i):

$$
\begin{aligned}
\mathbf{q}_i^{lbf} &= \left[\mathbf{W}_q^{lbf} \mathbf{h}_i^{bca}, \mathbf{W}_{qe}^{lbf} \mathbf{e}_i^{qb} \right], \\
\mathbf{k}_{ij}^{lbf} &= \left[\mathbf{W}_k^{lbf} \mathbf{h}_j^{bca} + \mathbf{W}_{pos}^{lbf} \mathbf{p}_{ij}, \mathbf{W}_{ke}^{lbf} \mathbf{e}_j^{qb} \right], \\
\mathbf{v}_{ij}^{lbf} &= \mathbf{W}_v^{lbf} \mathbf{h}_j^{bca},
\end{aligned}
\tag{7}
$$

where $\mathbf{h}_i^{bca}$ and $\mathbf{h}_j^{bca}$ are the features of future behaviors output from cross attention module for any target agent i and agent j in any scene modality, while $\mathbf{e}_i^{qb}$, $\mathbf{e}_j^{qb}$ are the corresponding behavior embedding, respectively. To make $\mathbf{e}_i^{qb}$ aware of the position of $\mathbf{e}_j^{qb}$, the location of any future trajectory of agent j is normalized to the polyline coordinate of agent i before embedding. The resulting query, key and value are utilized for low-level behavior fusion (lbf) using Eq. 2 and Eq. 3, where the neighbor $\mathcal{N}_i$ contains all future behavior features in the current scene modality from other target agents. Finally, we obtain the updated feature $\mathbf{h}_i^{lbf} \in \mathbb{R}^d$ for the future behavior of any target agent i in any scene modality, and the feature is used as the initialized query content for the next decoder layer.

Multi-modal decoder in LFBF. For each LFBF-based decoder layer, we append an MLP block to the output feature $\mathbf{h}_i^{lbf}$ for predicting future trajectory of any target agent i in any scene modality.

3.5. Training Loss

The training loss is composed of goal regression L_G and trajectory regression L_T for the total A target agents. For the multiple scene modalities, a winner-takes-all (WTA) strategy is taken to address mode collapse. More specifically, both two losses are only calculated for the scene modality that has the minimum final displacement error summed by all target agents, e.g., the sum of Euclidean distance between the endpoint of predicted trajectory (output

from the final decoder layer in LFBF) and ground truth:

$$L_G = \sum_{a=1}^{A} L_{reg}(g^a, \hat{g}_{k*}^a), \quad L_T = \sum_{a=1}^{A} L_{reg}(\mathbf{y}^a, \hat{\mathbf{y}}_{k*}^a) \quad (8)$$

where g^a and $\hat{g}_{k*}^a$ are the GT endpoint and regressed goal from HFIF decoder, respectively. $\mathbf{y}^a$ and $\hat{\mathbf{y}}_{k*}^a$ are the GT trajectory and predicted trajectory from LFBF-based decoder layer, respectively. L_{reg} is the smooth $l1$ loss. k^* is the index of the selected scene modality. The final loss is the sum of the L_G and the L_T with equal weights, where the L_T is applied to each LFBF-based decoder layer.

For inference, trajectories generated by the last decoder layer are considered as the final prediction result. The joint likelihood of each scene modality is calculated by the product of all target agent scores in each modality. To obtain agent modality score, we take the endpoint of each predicted trajectory to query the closest intention in the corresponding marginal heatmap (conditional anchors), then the queried intention score is used to calculate joint likelihoods.

4. Experiments

4.1. Experimental Setup

Dataset and metrics. We train and evaluate our BiFF in the Waymo Open Motion Dataset (WOMD) [11], a large-scale dataset with the most diverse interactive cases collected from realistic traffic scenes. We focus on the interactive challenge where the task is to predict 8s future trajectories for two interactive agents given 1s tracking history. The WOMD contains about 487k training samples, 44k validation samples and 44k testing samples. The metrics [11] are calculated with the official tool. Besides, to measure the consistency of the joint trajectory prediction for all K scene modalities, we adopt Cross Collision Rate (CCR) [2] [48] metric. Concretely, for a modality of a case, the collision is counted if the two agents overlap at any predicted time step. Then the ratio of collision is calculated by N_{coll}/K. The final CCR is averaged over all samples.

Implementation Details. Our model is trained for 30 epochs with 4 RTX 3090 GPUs using AdamW optimizer, and the initial learning rate, batch size are set to 0.0001 and 80 respectively. The learning rate is half decayed every 2 epochs from epoch 20. For the scene encoder, we stack $N_E = 6$ layers for the transformer encoder. The map is pruned with marginal heatmaps using half circle and half ellipse to ensure model accurately focusing on the interest area. The number of neighbors k is set to 16 for local self-attention. The hidden feature d is set as 256. For the

---

[2]Note that the official metric Overlap Rate (OR) in the challenge only calculates collision between prediction and ground truth for the modality with the highest score but not count overlap among joint predicted trajectories and not including other scene modalities
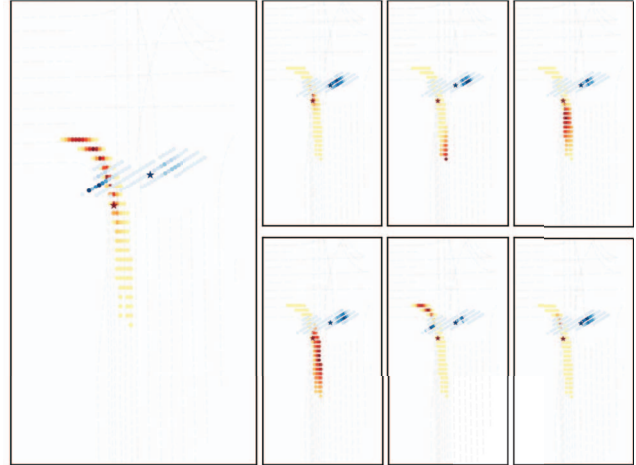

Figure 6. Illustration of conflict resolution in HFIF model. **Left**: Marginal heatmaps. **Right**: Scene-compliant assignment scores from different headers of motion decoder. Brightness in red and blue signifies the score of two interacting agents, with the asterisk denoting the ground truth goal points.

transformer decoder, we stack $N_H = 1$ layer for HFIF and $N_L = 3$ layers for LFBF. The number of map polylines L is set to 256 for scene context aggregation. The number of static future intentions S is set to 100 as the conditional anchors for HFIF. No data augmentation and model ensemble is used. More details can be referred to appendix.

4.2. Comparison with the State-of-the-art

We compare our BiFF with the SOTA methods on both validation and test interactive WOMD. The quantitative results of proposed BiFF and all baselines are shown in Tab. 1, where BiFF-200 denotes BiFF with intention number $S = 200$. For the results in validation set, our BiFF outperforms both Waymo Full Baseline and Scene Transformer in terms of all metrics except for mAP over cyclist. For the results in test set, compared to the approaches without ensemble models, BiFF achieves SOTA in minFDE and MR. Compared to the approaches with ensemble models, BiFF still achieves SOTA in minFDE over vehicle and pedestrian while all other metrics are near SOTA except for mAP. Although BiFF has lower mAP, it improves the minFDE and MR by a large margin over the methods without ensemble models, meaning BiFF predicts more accurate trajectories with efficient scene multi-modality. To illustrate that, we visualize the representative qualitative results in Fig. 9.

4.3. Ablation Study

To study the importance of each module in BiFF, we conduct ablation studies on the Waymo interactive validation set over joint metrics of vehicles at 8s. To improve the efficiency of ablation study, we resort to the alternative training

Table 1. Comparison with top-ranked entries on the WOMD Interaction Leaderboard. † refer to ensemble methods.

	Method	Vehicle(8s)			Pedestrain(8s)			Cyclist(8s)		
		mFDE↓	MR↓	mAP↑	mFDE↓	MR↓	mAP↑	mFDE↓	MR↓	mAP↑
Test	Waymo LSTM baseline [11]	12.40	0.87	0.01	6.85	0.92	0.00	10.84	0.97	0.00
	HeatIRm4 [29]	7.20	0.80	0.07	4.06	0.80	0.05	6.69	0.85	0.01
	AIR^2 [45]	5.00	0.64	0.10	3.68	0.71	0.04	5.47	0.81	0.04
	SceneTransformer [31]	4.08	0.50	0.10	3.19	0.62	0.05	4.65	0.70	0.04
	DenseTNT [17]	4.75	0.52	**0.20**	3.32	0.59	**0.11**	5.15	0.73	**0.05**
	M2I [41]	5.65	0.57	0.16	3.73	0.60	0.06	6.16	0.74	0.03
	BiFF (Ours)	3.82	**0.46**	0.13	2.79	**0.54**	0.06	4.33	**0.68**	0.03
	BiFF-200 (Ours)	**3.71**	0.47	0.12	**2.73**	0.56	0.05	**4.29**	0.69	0.03
	† MTR [38]	4.04	0.49	0.21	2.86	0.53	0.13	4.24	0.65	0.05
	† JFP [27]	3.88	0.45	0.19	2.81	0.54	0.15	4.19	0.63	0.06
Val	Waymo Full Baseline [11]	6.07	0.66	0.08	4.20	1.00	0.00	6.46	0.83	0.01
	SceneTransformer [31]	3.99	0.49	0.11	3.15	0.62	0.06	4.69	0.71	**0.04**
	BiFF (Ours)	3.74	**0.44**	**0.16**	2.70	**0.53**	0.06	4.42	**0.67**	0.03
	BiFF-200 (Ours)	**3.64**	0.46	**0.16**	**2.64**	0.54	**0.07**	**4.31**	0.69	0.03

scheme with total 15 epochs where the learning rate is initialized as 0.0001 and decayed to half from epoch 10.

Importance of HFIF. For simplicity, we name Goal Regression (GR) indicating the transformer decoder in Sec.3.3 without HFIF. For the HFIF-based transformer decoder, illustrated in Tab. 2, with HFIF module, MR is noticeably improved, which indicates that in interactive scenarios, HFIF plays a crucial role in goal prediction since the future intentions of target agents are likely to be intensely interactive. The future intentions of the interactive counterpart provide useful future information for predicting goals of each target.

In Fig. 6, an interactive sample is visualized to show the scene-consistent goals generated from multi-modal decoder in HFIF. On the left, the marginal heatmaps (conditional anchors) of two interactive agents are highly multi-modal and conflicting since the peaks are overlapped with each other. Our HFIF is aimed to decompose multi-modal to uni-modal [49] for each heatmap while splitting goals across target agents in each scene modality. On the right, we show the assignment scores of both two interactive agents predicted by K different headers. We observe that each header focuses on the distinctive interactive modality and the overlap is successfully eliminated in each scene modality, which verifies that the high-level future intention interaction across target agents is modeled efficiently by our HFIF.

Table 2. Effects of different components in BiFF. **GR**: Goal Regression, **TR**: Trajectory Refine, **CCR**: Cross Collsion Rate.

GR	HFIF	TR	LFBF	mADE↓	mFDE↓	MR↓	CCR↓
✓	✗	✗	✗	2.2292	5.0402	0.6033	0.1932
✓	✓	✗	✗	2.0167	4.5718	0.5572	0.1810
✓	✓	✓	✗	1.9283	4.3520	0.5564	0.1547
✓	✓	✓	✓	**1.8401**	**4.1497**	**0.5121**	**0.1397**

Importance of LFBF. For simplicity, the transformer decoder in Sec.3.4 without LFBF is termed as Trajectory

Refine (TR). From Tab. 2, for the LFBF-based transformer decoder, both components TR and LFBF can improve the performance. First, with TR module, the model can iteratively refine the future trajectory by aggregating the behavior-specific scene context in the future. We find that MR is not improved compared with HFIF, which further verifies the efficiency of HFIF in predicting goals. Moreover, the CCR significantly improves with TR, presumably because the predicted future behavior is more similar to the realistic driving scenario using TR. Thus, TR makes it more reliable for LFBF to perform local future fusion. Second, with LFBF module, the prediction of each target agent can rely on the future location of the interactive counterpart, which is useful since reliable information is strongly associated with future predictions. All metrics are continuously improved, which suggests the importance of LFBF in fusing low-level future behavior across interactive agents.

Importance of polyline-based coordinate. In order to substantiate the efficacy of polyline-based coordinates in enhancing data efficiency and frame robustness, we design a variant of BiFF with scene-centric reference frame, where all scene context is translated to the center location of two target agents at the current time step. To demonstrate data efficiency, we train BiFF and scene-centric BiFF using 5%, 10%, 20% and 100% frames uniformly sampled from the WOMD training set, and evaluate them on the interactive validation set. From Fig. 7, we can find that compared to scene-centric variant, BiFF achieves the same performance with fewer training data. The final prediction accuracy also becomes better, which verifies the improved ability of model generalization with polyline-based coordinate. To show the robustness to coordinate, from Fig. 8, we can observe that our BiFF is robust to the rotation while the performance of the variant severely gets worse with the angle variation. Besides, by normalizing scene-centric BiFF to
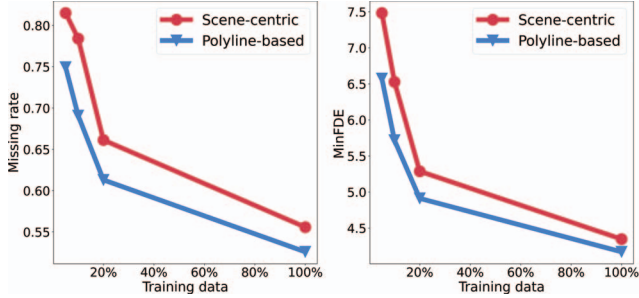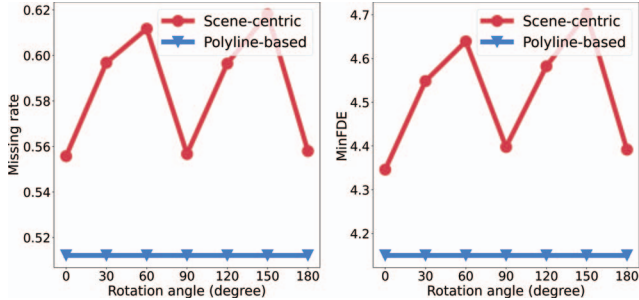
Figure 7. Data efficiency.

Figure 8. Robustness to rotation angles.

a. Yield for pedestrian   b. Yield for high-speed vehicle

c. Long-term interactive merging   d. Wait for the red light

Figure 9. Qualitative results on WOMD validation set (K=3).

any polyline-based coordinate, the results become not applicable but the performance of our BiFF is still unchanging, which verifies the robustness to the variant of reference frame by adopting polyline-based coordinate.

We measure memory efficiency with the Average Polyline Number (APN) for each target agent on the Waymo interactive validation set. By using polyline-based coordinate, the agent-centric strategy is achieved without redundant context, all target agents in each case share the scene context and the APN is 179. However, when encoding local context for each target separately, it causes the redundant context and the APN is 291 meaning requiring much more memory during training and inference.

Table 3. Effects of quantity of conditional anchors.

Anchor Num	mADE↓	mFDE↓	MR↓
50	1.9442	4.4522	0.5229
100	1.8401	4.1497	**0.5121**
200	**1.8284**	**4.1058**	0.5434
500	1.9405	4.4488	0.6009

Table 4. Effects of quality of conditional anchors.

Anchor Quality (mFDE)	mADE↓	mFDE↓	MR↓
4.24	2.1138	4.9617	0.6371
3.99	1.9572	4.4918	0.5554
3.42 **w/o** score	**1.8401**	**4.1497**	**0.5121**
3.42 **w** score	1.8501	4.1694	0.5258

Analysis of conditional anchors. We analyze the effects of the quantity and quality of conditional anchors. For the quantity of conditional anchors, by changing the num-
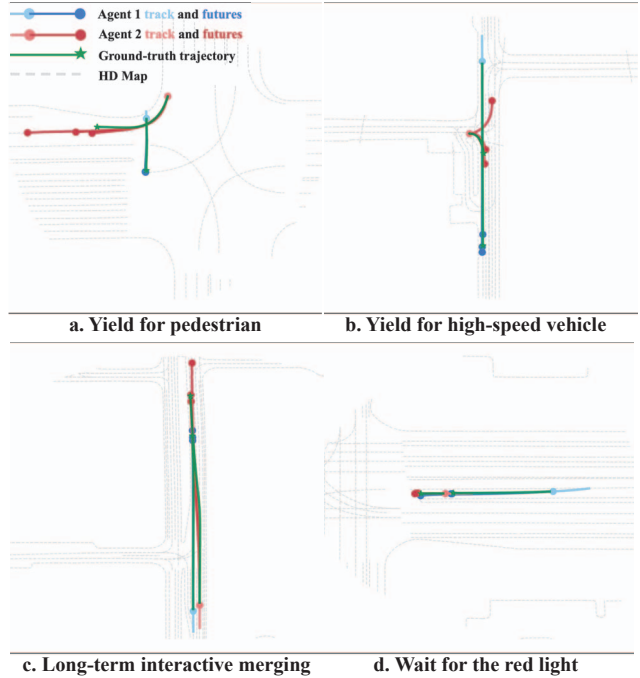
ber of intentions in Tab. 3, we observe that within a certain range, the minFDE is improved with anchor number increasing while the minMR becomes worse. For the quality of conditional anchors, we define marginal minFDE [3] to quantify the quality of marginal heatmap (conditional anchors). From Tab. 4, all metrics are improved with quality enhancement, which motivates the prediction performance to be further improved by using higher-quality marginal heatmaps. We also find that the performance is slightly dropped by adding the confidence score to each intention, and the marginal score may mislead the model.

5. Conclusion

In this paper, we propose BiFF to model the future interactions, where HFIF fuses high-level future intentions and LFBF fuses low-level future behaviors. With polyline-based coordinates, BiFF becomes data-efficient, robust to global frame and more accurate in prediction. Experimental results demonstrate that BiFF achieves SOTA performance, underscoring the importance of each proposed module.

---

[3] We obtain 6 endpoints by using NMS on the marginal heatmap, and marginal minFDE is obtained by evaluating on the validation set.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–971, 2016. 1, 2

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020. 1, 2

[3] Sergio Casas, Cole Gulino, Simon Suo, Katie Luo, Renjie Liao, and Raquel Urtasun. Implicit latent variable model for scene-consistent motion forecasting. In *Proc. European Conference on Computer Vision (ECCV)*. Springer, 2020. 1, 2

[4] Yuning Chai, Benjamin Sapp, Mayank Bansal, and Dragomir Anguelov. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. In *Proc. Conference on Robot Learning (CoRL)*, 2019. 1, 2

[5] Xiaobo Chen, Huanjia Zhang, Feng Zhao, Yu Hu, Chenkai Tan, and Jian Yang. Intention-aware vehicle trajectory prediction based on spatial-temporal dynamic attention network for internet of vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):19471–19483, 2022. 2

[6] Alexander Cui, Sergio Casas, Kelvin Wong, Simon Suo, and Raquel Urtasun. Gorela: Go relative for viewpoint-invariant motion forecasting. *arXiv preprint arXiv:2211.02545*, 2022. 2

[7] Henggang Cui, Vladan Radosavljevic, Fang-Chieh Chou, Tsung-Han Lin, Thi Nguyen, Tzu-Kuo Huang, Jeff Schneider, and Nemanja Djuric. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *Proc. IEEE Int. Conference on Robotics and Automation (ICRA)*, pages 2090–2096. IEEE, 2019. 1, 2

[8] Nachiket Deo and Mohan M Trivedi. Convolutional social pooling for vehicle trajectory prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1468–1476, 2018. 1, 2

[9] Wenchao Ding, Jing Chen, and Shaojie Shen. Predicting vehicle behaviors over an extended horizon using behavior interaction network. In *2019 international conference on robotics and automation (ICRA)*, pages 8634–8640. IEEE, 2019. 1

[10] Nemanja Djuric, Vladan Radosavljevic, Henggang Cui, Thi Nguyen, Fang-Chieh Chou, Tsung-Han Lin, Nitin Singh, and Jeff Schneider. Uncertainty-aware short-term motion prediction of traffic actors for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2095–2104, 2020. 2

[11] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021. 1, 2, 7, 8

[12] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11525–11533, 2020. 1, 2, 3

[13] Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Home: Heatmap output for future motion estimation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 500–507. IEEE, 2021. 2

[14] Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Thomas: Trajectory heatmap output with learned multi-agent sampling. *arXiv preprint arXiv:2110.06607*, 2021. 1, 2, 3

[15] Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Gohome: Graph-oriented heatmap output for future motion estimation. In *2022 international conference on robotics and automation (ICRA)*, pages 9107–9114. IEEE, 2022. 2

[16] Roger Girgis, Florian Golemo, Felipe Codevilla, Martin Weiss, Jim Aldon D'Souza, Samira Ebrahimi Kahou, Felix Heide, and Christopher Pal. Latent variable sequential set transformers for joint multi-agent motion prediction. In *International Conference on Learning Representations*, 2022. 1, 2, 3

[17] Junru Gu, Chen Sun, and Hang Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15303–15312, 2021. 1, 2, 8

[18] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2255–2264, 2018. 2

[19] Dirk Helbing and Peter Molnar. Social force model for pedestrian dynamics. *Physical review E*, 51(5):4282, 1995. 2

[20] Joey Hong, Benjamin Sapp, and James Philbin. Rules of the road: Predicting driving behavior with a convolutional model of semantic interactions. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8454–8462, 2019. 2

[21] Xiaosong Jia, Liting Sun, Hang Zhao, Masayoshi Tomizuka, and Wei Zhan. Multi-agent trajectory prediction by combining egocentric and allocentric views. In *Conference on Robot Learning*, pages 1434–1443. PMLR, 2022. 2

[22] Namhoon Lee, Wongun Choi, Paul Vernaza, Christopher B Choy, Philip HS Torr, and Manmohan Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 336–345, 2017. 1, 2

[23] Jiachen Li, Fan Yang, Masayoshi Tomizuka, and Chiho Choi. Evolvegraph: Multi-agent trajectory prediction with dynamic relational reasoning. *Advances in neural information processing systems*, 33:19783–19794, 2020. 2

[24] Xin Li, Xiaowen Ying, and Mooi Choo Chuah. Grip: Graph-based interaction-aware trajectory prediction. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 3960–3966. IEEE, 2019. 2

[25] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 541–556. Springer, 2020. 1, 2

[26] Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. Multimodal motion prediction with stacked transformers. *Computer Vision and Pattern Recognition*, 2021. 2

[27] Wenjie Luo, Cheolho Park, Andre Cornman, Benjamin Sapp, and Dragomir Anguelov. Jfp: Joint future prediction with interactive multi-agent modeling for autonomous driving. *arXiv preprint arXiv:2212.08710*, 2022. 2, 8

[28] Jean Mercat, Thomas Gilles, Nicole El Zoghby, Guillaume Sandou, Dominique Beauvois, and Guillermo Pita Gil. Multi-head attention for multi-modal joint vehicle motion forecasting. In *Proc. IEEE Int. Conference on Robotics and Automation (ICRA)*, pages 9638–9644. IEEE, 2020. 2

[29] Xiaoyu Mo, Zhiyu Huang, and Chen Lv. Multi-modal interactive agent trajectory prediction using heterogeneous edge-enhanced graph attention network. In *Workshop on Autonomous Driving, CVPR*, volume 6, page 7, 2021. 8

[30] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14424–14432, 2020. 2

[31] Jiquan Ngiam, Vijay Vasudevan, Benjamin Caine, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, et al. Scene transformer: A unified architecture for predicting future trajectories of multiple agents. In *International Conference on Learning Representations*, 2022. 1, 2, 3, 8

[32] Tung Phan-Minh, Elena Corina Grigore, Freddy A Boulton, Oscar Beijbom, and Eric M Wolff. Covernet: Multimodal behavior prediction using trajectory sets. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14074–14083, 2020. 2

[33] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 3, 6

[34] Nicholas Rhinehart, Kris M Kitani, and Paul Vernaza. R2p2: A reparameterized pushforward policy for diverse, precise generative path forecasting. In *Proc. European Conference on Computer Vision (ECCV)*, pages 772–788, 2018. 2

[35] Nicholas Rhinehart, Rowan McAllister, Kris Kitani, and Sergey Levine. Precog: Prediction conditioned on goals in visual multi-agent settings. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 2821–2830, 2019. 2

[36] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1349–1358, 2019. 2

[37] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Proc. European Conference on Computer Vision (ECCV)*, 2020. 2

[38] 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. 1, 2, 4, 6, 8

[39] Haoran Song, Wenchao Ding, Yuxuan Chen, Shaojie Shen, Michael Yu Wang, and Qifeng Chen. Pip: Planning-informed trajectory prediction for autonomous driving. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 598–614. Springer, 2020. 2

[40] Haoran Song, Di Luan, Wenchao Ding, Michael Y Wang, and Qifeng Chen. Learning to predict vehicle trajectories with model-based planning. In *Conference on Robot Learning*, pages 1035–1045. PMLR, 2022. 2

[41] 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. 1, 2, 8

[42] Yichuan Charlie Tang and Ruslan Salakhutdinov. Multiple futures prediction. In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, 2019. 2

[43] Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen, Bertrand Douillard, Chi Pang Lam, Dragomir Anguelov, et al. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7814–7821. IEEE, 2022. 2

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 2, 4

[45] David Wu and Yunnan Wu. AIR^2 for interaction prediction. *arXiv preprint arXiv:2111.08184*, 2021. 8

[46] Maosheng Ye, Tongyi Cao, and Qifeng Chen. Tpcn: Temporal point cloud networks for motion forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11318–11327, 2021. 2

[47] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9813–9823, 2021. 2

[48] Wei Zhan, Liting Sun, Di Wang, Haojie Shi, Aubrey Clausse, Maximilian Naumann, Julius Kummerle, Hendrik Konigshof, Christoph Stiller, Arnaud de La Fortelle, et al. Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with

semantic maps. *arXiv preprint arXiv:1910.03088*, 2019. 2, 7

[49] Lu Zhang, Peiliang Li, Jing Chen, and Shaojie Shen. Trajectory prediction with graph-based dual-scale context fusion. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11374–11381. IEEE, 2022. 8

[50] Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Benjamin Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. *Proc. Conference on Robot Learning (CoRL)*, 2020. 1, 2

[51] Tianyang Zhao, Yifei Xu, Mathew Monfort, Wongun Choi, Chris Baker, Yibiao Zhao, Yizhou Wang, and Ying Nian Wu. Multi-agent tensor fusion for contextual trajectory prediction. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12126–12134, 2019. 2

[52] Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8833, 2022. 2