IMPACT: 2nd Place Solution for 2025 Waymo Open Dataset Challenge - Interaction Prediction

Jiawei Sun[1]*, Xibin Yue[2], Jiahui Li[1], Tianle Shen[1], Chengran Yuan[1],
Shuo Sun[1], Sheng Guo[1], Quanyun Zhou[2], Marcelo H. Ang Jr[1]
[1]National University of Singapore [2]Xiaomi EV

Abstract

We present IMPACT, a unified trajectory prediction framework that jointly models both behavioral intentions and future trajectories, aiming to enhance prediction accuracy, interpretability, and efficiency. After a shared context encoding stage, IMPACT simultaneously predicts agent-level behavioral intentions and polyline-level occupancy within a symmetric architecture, providing strong priors for trajectory decoding through dynamic, modality-dependent context pruning. Notably, our single-model performance achieves second place in the Interaction Prediction track of the 2025 Waymo Open Dataset Challenge.

1. Introduction

Trajectory prediction in interactive driving scenarios is one of the most challenging problems in autonomous driving. To enable safe and social-aware driving, a prediction model must not only forecast the future motions of surrounding agents, such as vehicles, pedestrians, and cyclists, but also capture the underlying interactions among them. Most prior multi-agent prediction methods estimate marginal distributions for each agent, assuming conditional independence. This overlooks future social interactions that are critical for accurate scene understanding and coordinated decision-making. Most existing methods determine attention inputs purely based on geometric distance, typically selecting nearby agents and map elements through fixed-radius [1] or k-nearest-neighbor (KNN) filtering [2]. While effective in reducing computation, these distance-based filters are oblivious to semantic intent and interaction relevance, often leading to the exclusion of critical context or the inclusion of irrelevant elements.

In this report, we introduce **IMPACT**, an intention-aware trajectory prediction framework. Our architecture begins with a shared encoder that extracts rich contextual information from both agents and map elements. This encoded representation is then jointly leveraged by three core modules: an agent-level behavioral intention predictor, a polyline-level vectorized occupancy predictor, and a trajectory decoder. The predicted intentions and occupancy probabilities serve as priors, dynamically filtering out irrelevant agents and map elements for trajectory decoding stage. This design enables IMPACT to produce accurate, interpretable, and computationally efficient multimodal trajectory predictions.

*This work was done during his internship at Xiaomi EV.

2. Methodologies

2.1. Input Representation

In our method, we apply agent-centric normalization. To predict a target agent, the input to the predictor comprises: $A = \{a_1, a_2, \ldots, a_{N_a}\} \in \mathbb{R}^{N_a \times T_p \times F_1}$, representing N_a agents with T_p past states and feature dimension F_1, and $\mathcal{L} = \{l_1, l_2, \ldots, l_{N_l}\} \in \mathbb{R}^{N_l \times N_p \times F_2}$, representing N_l polylines with N_p points each and feature dimension F_2.

Following the approaches of RMP-YOLO[3], we further incorporate the historical relative movement between the target agent and the map polylines to capture dynamic subtle interdependencies. This historical movement is denoted by $\mathcal{R} = \{r_1, r_2, \ldots, r_{N_l}\} \in \mathbb{R}^{N_l \times T_p \times F_3}$, where F_3 is the feature size associated with the relative movement($(\Delta x, \Delta y, \cos\Delta\theta, \sin\Delta\theta)$).

2.2. Network Structure

2.2.1. Spatial Temporal Encoding

To comprehensively model temporal dependencies, we apply a Multi-Scale LSTM (MSL) module. The time-series data $\mathcal{A}$ and $\mathcal{R}$ are each processed through three parallel streams. Each stream consists of a 1D CNN with a distinct kernel size followed by an LSTM. For the i-th stream with kernel size k_i, the output at the final time step T_p is computed as:

$$\mathcal{A}_{k_i}^{T_p} = \text{LSTM}\left(\text{Conv1D}_{k_i}(\mathcal{A})\right)\Big|_{t=T_p}, \quad k_i \in \{1,3,5\} \tag{1}$$

$$\mathcal{R}_{k_i}^{T_p} = \text{LSTM}\left(\text{Conv1D}_{k_i}(\mathcal{R})\right)\Big|_{t=T_p}, \quad k_i \in \{1,3,5\} \tag{2}$$

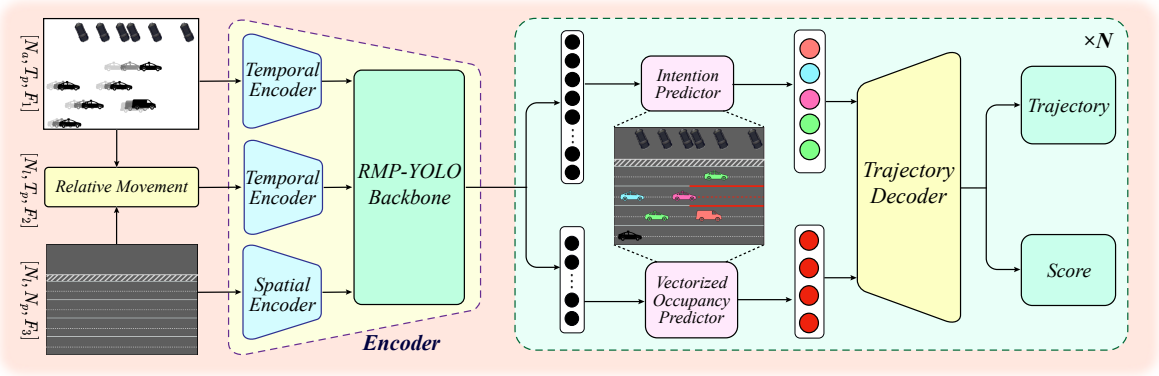

Figure 1. An overview of framework of IMPACT. Both the Intention Predictor and the Vectorized Occupancy Predictor share the same context encoder with the Trajectory Decoder, leveraging their outputs to prune irrelevant agents and map polylines. This selective mechanism ensures that only the most critical context is fed into the decoder for final trajectory prediction.

The final-step hidden states are concatenated ($\oplus$) along the feature dimension, and a multi-layer perceptron (MLP) projects the concatenated vector into a unified temporal feature token:

$$\mathcal{A}^1 = MLP(\mathcal{A}_1^{T_p} \oplus \mathcal{A}_3^{T_p} \oplus \mathcal{A}_5^{T_p}) \in \mathbb{R}^{N_a \times D}, \tag{3}$$

$$\mathcal{R}^1 = MLP(\mathcal{R}_1^{T_p} \oplus \mathcal{R}_3^{T_p} \oplus \mathcal{R}_5^{T_p}) \in \mathbb{R}^{N_l \times D}. \tag{4}$$

For the spatial data $\mathcal{L}$, we adopt a simplified PointNet-like architecture to aggregate each polyline into a feature token:

$$\mathcal{L}_1 = \text{MaxPooling}(\text{MLP}(\mathcal{L})) \in \mathbb{R}^{N_l \times D}. \tag{5}$$

These tokens are aligned in the feature space ($\mathbb{R}^D$) and ready for downstream fusion and prediction tasks.

2.2.2. Feature Fusion.

We adopt the encoder from RMP-YOLO [3] as the backbone network for feature fusion (see Figure 1). To integrate heterogeneous input modalities, we employ a cascaded Multi-Context Gating (MCG) mechanism inspired by [4]. The MCG modules sequentially fuse pairs of modalities from a candidate set of three. The output of each MCG stage serves as input to the subsequent stage, enabling hierarchical feature interaction:

$$(\mathcal{A}^2, \mathcal{R}^2) = \text{MCG}(\mathcal{A}^1, \mathcal{R}^1), \tag{6}$$

$$(\mathcal{L}^2, \mathcal{R}^3) = \text{MCG}(\mathcal{L}^1, \mathcal{R}^2), \tag{7}$$

$$(\mathcal{A}^3, \mathcal{L}^3) = \text{MCG}(\mathcal{A}^2, \mathcal{L}^2), \tag{8}$$

The final fused tokens are defined as agent tokens: $\mathcal{A}_3 \in \mathbb{R}^{N_a \times D}$ and map tokens: $\mathcal{L}_3 = \mathcal{L}_3 + \mathcal{R}_3 \in \mathbb{R}^{N_p \times D}$. Thus, we design a K-nearest-neighbor (KNN) guided local attention mechanism to restrict each agent token to attend only to its K most relevant neighboring tokens (agents or map elements). This sparse attention pattern reduces computational complexity while preserving critical interactions. Six transformer encoder layers are applied to achieve deep feature fusion. Each layer follows the standard transformer architecture enhanced with positional encoding:

$$\begin{aligned} X^i = \text{MHA}(&X^{i-1} + \text{PE}(X^{i-1}), \\ &\mathcal{K}(X^{i-1}) + \text{PE}(\mathcal{K}(X^{i-1})), \mathcal{K}(X^{i-1})) \end{aligned} \tag{9}$$

where $X^0 = [\mathcal{L}_3, \mathcal{A}_3] \in \mathbb{R}^{(N_a+N_m)\times D}$, $\text{MHA}(\cdot)$ denotes multi-head attention, $\mathcal{K}(\cdot)$ selects K-nearest neighbors via Euclidean distance, and $\text{PE}(\cdot)$ injects positional information using sinusoidal encoding. The positional coordinates derive from agents' latest observed positions and map polylines' centroid coordinates. The final output tokens $X^{Final} = [\mathcal{L}_4, \mathcal{A}_4]$ are fed into the behavioral intention prediction module and vectorized occupation prediction for dynamic context-aware pruning.

Before diving into decoder part (see Figure 2), we define query content feature at decoder layer i as $Q^i \in \mathbb{R}^{K\times D}$, which are later used to aggregate information from agent tokens and map tokens, and decode multimodal prediction results and K denotes the number of different futures.

2.2.3. Multimodal Behavioral Intention Prediction

For each future modality, we predict the behavioral intentions of other agents with respect to the target agent. Given the input agent tokens $\mathcal{A}_4 \in \mathbb{R}^{N_a\times D}$ and the query content $Q^i \in \mathbb{R}^{K\times D}$, we fuse these features into a unified representation of shape $\mathbb{R}^{K\times N_a\times 2D}$ via straightforward tensor broadcasting and concatenation. Next, the fused features are passed through a multi-layer perceptron (MLP) and then added to the previous layer's behavioral intention token $I^{i-1} \in \mathbb{R}^{K\times N_a\times D}$. Finally, another MLP followed by a softmax activation function produces the final behavioral intention predictions:

$$\begin{aligned} \hat{\text{H}}^i &= \text{Softmax}\Big(\text{MLP}\big(I^i\big)\Big) \in \mathbb{R}^{K\times N_a\times 4}, \\ I^i &= \text{MLP}(\text{MLP}(\mathcal{A}_4 \oplus Q^i) + I^{i-1}). \end{aligned} \tag{10}$$

Each vector element h represents a probability distribution over four intention categories: *yielding*, *overtaking*, *ignored*, and *nearby*. To make the decoding process

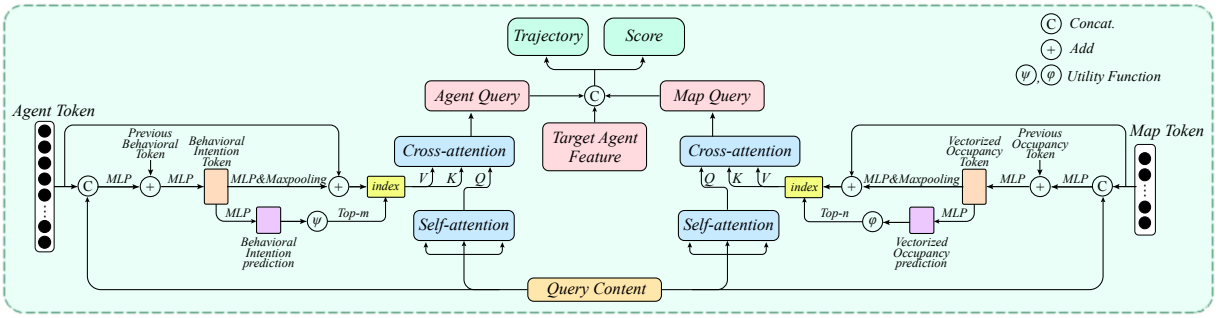

Figure 2. An overview of our decoder framework, featuring context-aware pruning via symmetric dual filters.

more focused, we first compute an overall interaction score from the predicted distribution $\hat{\mathrm{H}}$ using a utility function ψ, producing $\psi(\hat{\mathrm{H}}^i) \in \mathbb{R}^{K \times N_a}$. We then select the top-m highest-scoring agents for downstream trajectory decoding. Therefore, for each modality, we choose m most revelant agents $\mathcal{A}_5 \in \mathbb{R}^{m \times D}$, where $m \ll N_a$. This filtering step refines the decoder's input, concentrating on the most influential interactions while improving prediction accuracy. The ground-truth label of behavioral intention H^* is derived from an auto-labeled data preprocessing process.

2.2.4. Multimodal Vectorized Occupancy Prediction

Unlike conventional occupancy prediction methods that rely on computationally intensive rasterization of multi-view images, we introduce a novel vectorized occupancy prediction framework that integrates seamlessly with our vectorized scenario representation. For each map polyline l_i, we predict multimodal occupancy probabilities corresponding to different future hypotheses of the target agent. Denoting C^{i-1} as the previous vectorized occupancy tokens, we apply an operation symmetric to the multimodal behavioral intention prediction:

$$\begin{aligned} \hat{\mathrm{O}}^i &= \mathrm{Sigmoid}\Big(\mathrm{MLP}\big(C^i\big)\Big) \in \mathbb{R}^{K \times N_l \times 1}, \\ C^i &= \mathrm{MLP}(\mathrm{MLP}(\mathcal{L}_4 \oplus Q^i) + C^{i-1}) \end{aligned} \tag{11}$$

This vectorized approach ensures both efficiency and scalability while maintaining alignment with the overall vectorized representation of the scene. Among the N_l polylines's multimodal occupancy probabilities $\varphi(\hat{\mathrm{O}}^i) \in \mathbb{R}^{K \times N_l}$, we select the top-$n$ with the highest predicted probabilities in each modality to form $\mathcal{L}_5 \in \mathbb{R}^{n \times D}$, where $n \ll N_l$. These top-ranked polylines serve as focused inputs for the subsequent trajectory decoder. The ground-truth occupancy label O^* is also derived from an auto-labeled data preprocessing process.

2.2.5. Trajectory Decoder

We adopt a multi-layer MTR-style[2] trajectory decoder. At each layer i, self-attention is applied to the query content $Q^i \in \mathbb{R}^{K \times D}$ across the K motion modes, enabling information exchange among different future modalities. Subsequently, for each modality, two cross-attention modules integrate features from the filtered agent tokens $\mathcal{A}_5$ and map tokens $\mathcal{L}_5$. Finally, the target agent feature (replicated K times) is concatenated with the cross-attended query features, and passed through a regression head to generate a set of Gaussian Mixture Model (GMM) parameters at each future timestep: $\left\{\left(\mu_x^k, \mu_y^k, \sigma_x^k, \sigma_y^k, \rho^k\right)\right\}_{k=1}^{K}$, where $\left(\mu_x^k, \mu_y^k, \sigma_x^k, \sigma_y^k, \rho^k\right)$ parameterizes the k-th Gaussian component. In addition, a classification head outputs the confidence scores $S \in \mathbb{R}^K$ corresponding to each motion mode. This multimodal representation captures the inherent uncertainties of agent trajectories.

2.3. Training Loss

Our overall training objective comprises four components:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{Int}} + \lambda_2 \mathcal{L}_{\text{Occ}} + \lambda_3 \mathcal{L}_{\text{Traj}} + \lambda_4 \mathcal{L}_{\text{Score}}, \tag{12}$$

where $\lambda_1, \lambda_2, \lambda_3$, and λ_4 are weighting factors balancing the contributions of behavioral intention prediction, vectorized occupancy prediction, trajectory prediction, and mode classification, respectively. Specifically, $\mathcal{L}_{\text{Int}}$ is calculated using the multi-class Focal Loss, $\mathcal{L}_{\text{Occ}}$ is based on the binary Focal Loss, $\mathcal{L}_{\text{Traj}}$ is derived from the GMM loss, and $\mathcal{L}_{\text{Score}}$ is computed with Binary Cross-Entropy. During training, the winner-take-all strategy is applied for $\mathcal{L}_{\text{Int}}$, $\mathcal{L}_{\text{Occ}}$, and $\mathcal{L}_{\text{Traj}}$, ensuring that only the modality closest to the ground-truth trajectory is used to compute these losses. The weighting factors are set as $\lambda_1 = 100$, $\lambda_2 = 100$, $\lambda_3 = 1$, and $\lambda_4 = 1$.

3. Experiments

3.1. Experimental Setup

Implementation Details. We employ AdamW optimizer for training, conducting experiments on a cluster of 8 NVIDIA A800 GPUs with a total batch size of 80. The learning rate is initialized as 1×10^{-4} and begins step decay starting at epoch 22, halving every two epochs. The model undergoes 30 epochs.

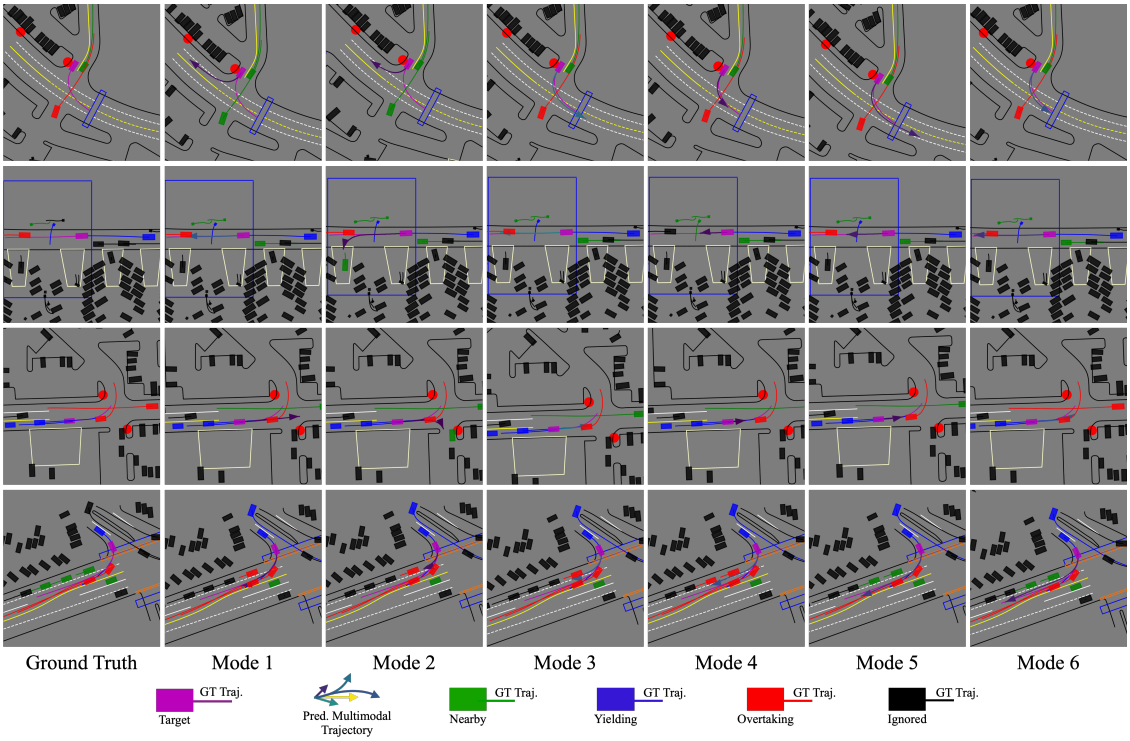

Figure 3. Visualization of predicted multi mode intention labels. The first column renders the ground truth intention labels of the agents. Remaining columns render the predicted results of K=6 different modes. We can see that our predicted intentions and trajectories are coupled and cover most possible modes.

Table 1. Leaderboard performance of the interaction prediction track of the Waymo Open Dataset Challenge.

Method	minADE ↓	minFDE ↓	Miss Rate ↓	mAP ↑	Soft mAP ↑
Parallel ModeSeq	0.7707	1.6897	0.3782	0.2949	0.2978
IMPACT(ours)[5]	0.9738	2.2734	0.4316	0.2659	0.2718
BeTop-ens [6]	0.9779	2.2805	0.4376	0.2511	0.2573
RetroMotion (SMoE hybrid)	0.9256	2.0890	0.4347	0.2519	0.2562
RMP_YOLO[3]	0.9274	2.1131	0.4167	0.2313	0.2486
AutoDiffuser-Draft	0.9422	2.1759	0.4885	0.2211	0.2249
infgen-base-xl	1.2059	2.6427	0.5649	0.0714	0.1236
Waymo LSTM baseline	1.9056	5.0278	0.7750	0.0648	0.0693

3.2. Leaderboard Performance

Joint Prediction Performance. As presented in Table 1, without any model-ensemble techniques, our single model achieves the second best performance on the Waymo joint prediction leaderboard. We visualize the per-mode qualitative results in Figure 3. Each predicted trajectory mode is conditioned on the behavioral intention outputs, which act as semantic priors for decoding. This confirms the effectiveness of intention-aware guidance in generating coherent and interaction-consistent predictions.

References

[1] Z. Zhou, J. Wang, Y.-H. Li, and Y.-K. Huang, "Query-centric trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 17 863–17 873. 1

[2] S. Shi, L. Jiang, D. Dai, and B. Schiele, "Motion transformer with global intention localization and local movement refinement," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 6531–6543. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/2ab47c960bfee4f86dfc362f26ad066a-Paper-Conference.pdf 1, 3

[3] J. Sun, J. Li, T. Liu, C. Yuan, S. Sun, Z. Huang, A. Wong, K. P. Tee, and M. H. A. Jr, "Rmp-yolo: A robust motion predictor for partially observable scenarios even if you only look once,"

2024. [Online]. Available: https://arxiv.org/abs/2409.11696 1, 2, 4

[4] B. Varadarajan, A. Hefny, A. Srivastava, K. S. Refaat, N. Nayakanti, A. Cornman, K. Chen, B. Douillard, C. P. Lam, D. Anguelov, and B. Sapp, "Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 7814–7821. 2

[5] J. Sun, X. Yue, J. Li, T. Shen, C. Yuan, S. Sun, S. Guo, Q. Zhou, and M. H. A. Jr, "Impact: Behavioral intention-aware multimodal trajectory prediction with adaptive context trimming," 2025. [Online]. Available: https://arxiv.org/abs/2504.09103 4

[6] H. Liu, L. Chen, Y. Qiao, C. Lv, and H. Li, "Reasoning multi-agent behavioral topology for interactive autonomous driving," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds., vol. 37. Curran Associates, Inc., 2024, pp. 92 605–92 637. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2024/file/a862f5788fd09bb6843c694d8120d50c-Paper-Conference.pdf 4