IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

IMPACT: Behavioral Intention-aware Multimodal Trajectory Prediction with Adaptive Context Trimming

***Abstract*—While most prior research has focused on improving the precision of multimodal trajectory predictions, the explicit modeling of multimodal behavioral intentions (e.g., yielding, overtaking) remains relatively underexplored. This paper proposes a unified framework that jointly predicts both behavioral intentions and trajectories to enhance prediction accuracy, interpretability, and efficiency. Specifically, we employ a shared context encoder for both intention and trajectory predictions, thereby reducing structural redundancy and information loss. Moreover, we address the lack of ground-truth behavioral intention labels in mainstream datasets (Waymo, Argoverse) by auto-labeling these datasets, thus advancing the community's efforts in this direction. We further introduce a vectorized occupancy prediction module that infers the probability of each map polyline being occupied by the target vehicle's future trajectory. By leveraging these intention and occupancy predictions priors, our method conducts dynamic, modality-dependent pruning of irrelevant agents and map polylines in the decoding stage, effectively reducing computational overhead and mitigating noise from non-critical elements. Our approach ranks first among LiDAR-free methods on the Waymo Motion Dataset and achieves SOTA performance on the Waymo Interactive Prediction Dataset. Remarkably, even without model ensembling, our single-model framework improves the softmAP by 10% compared to the previous SOTA method, BETOP, in Waymo Interactive Prediction Leaderboard. Furthermore, the proposed framework has been successfully deployed on real vehicles, demonstrating its practical effectiveness in real-world applications.**

I. Introduction

In the realm of autonomous driving, accurately predicting the future behaviors of surrounding agents is crucial to ensure safe, efficient, and comfortable driving experience.

With mainstream approaches including [1]–[4] prioritising generating precise multimodal trajectory predictions, most recent research focused on achieving higher average precisions (mAP) and minimizing deviations (Brier-minFDE). However, humans primarily determine the future behavior of interacting agents not by focusing on specific states alone, but rather by inferring the underlying intention that drives their actions.

This insight is also applicable in autonomous driving, where predicting surrounding agents' future poses alone is insufficient and partial; understanding their behavioral intention (eg, yielding or overtaking) toward the ego vehicle is equally critical. Without explicit intention information, the downstream decision-making and planning modules are left with only a distribution of possible future coordinates. They must then infer the other vehicle's underlying behavior in an additional step, which not only complicates the ego vehicle's decision-making process but can also reduce the overall reliability and responsiveness of the system. The most straightforward method is to additionally incorporate one more behavioral intention prediction module, but this decoupled architecture inevitably introduces structural redundancy and information fragmentation.

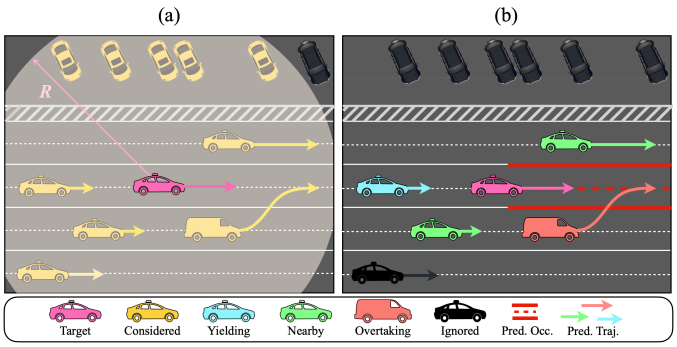

Fig. 1. (a) illustrates the traditional context input, while (b) is our integrated approach jointly predicting behavioral intentions, trajectories, and vectorized occupancy. In our approach, the decoder stage is fed only with influential agents and relevant map elements.

Currently, most trajectory prediction models use attention mechanism for querying information from agents and map features. However, indiscriminately attending to all agents and map elements introduces unnecessary complexity and potential causal confusion, as many elements have no direct relevance to the target agent, ultimately hindering correct decoding of the target agent's future trajectories. To address this, some recent approaches introduce an additional local context-aware refinement module after the decoder stage [5], [6], but this further complicates the prediction pipeline. Others adopt rule-based heuristic pruning of input lanes [7] or dynamically collect map elements based on the previous layer prediction results [3], [8]–[10]. However, these methods may struggle in complex scenarios or suffer from error propagation. BETOP [11] is the first paper explicitly modelling agent interactions through braid theory, but its topological selection algorithm demonstrates critical accuracy deficiencies , as illustrated in our supplementary video. Thus, an efficient approach to selectively focus on truly critical features (both agents and maps) remains an open challenge. Furthermore, the lack of ground-truth intention labels in current mainstream datasets (e.g., Waymo, Argoverse) is limiting the development of accurate behavioral intention-aware models.

To holistically address the aforementioned challenges, we propose our approach, IMPACT. Specifically, we introduce an agent-wise multimodal behavioral intention module before the trajectory decoder, while sharing the same scenario information extraction module (i.e., the same encoder) used by the trajectory predictor. This design reduces computational overhead and mitigates information loss. The agent-wise behavioral intention module outputs a one-hot intention vector for each vehicle, where each element in the vector represents the probability of a specific intention relative to the target vehicle (e.g., overtaking, yielding). Based on these probability vectors, we apply a utility function to assign an "interaction score"

to each vehicle, selecting only those with high interaction probabilities as inputs to the decoder. In parallel, we propose a vectorized occupancy prediction module. This module predicts the probability that each vectorized map element will be occupied by the target vehicle in the future. Similar to the agent selection process, we only feed the map polylines with high occupancy probabilities into the decoder. By integrating both modules in this manner, our method effectively reduces redundant interactions while preserving crucial information for accurate and robust trajectory prediction. We also propose an automatic labeling algorithm to generate high-quality ground truth labels for behavioral intentions on the Waymo and Argoverse 1&2 dataset.

We evaluated IMPACT on both the Waymo Motion Prediction Dataset (for marginal prediction) and the Waymo Interactive Prediction Dataset (for interaction prediction). Experimental results demonstrate that our method achieves the highest performance among all LiDAR-free approaches on the Waymo Motion Prediction Dataset, second only to MTRV3 [12], which benefits from additional LiDAR inputs. Furthermore, IMPACT sets a new state-of-the-art on the Waymo Interactive Prediction Dataset, surpassing the previous best-performing ensemble model [11], despite using only a single model. To further demonstrate the practical applicability of our approach, we successfully trained IMPACT on a proprietary dataset and deployed it effectively on a real vehicle.

Our main contributions can be summarized as follows:

1) **Intent-Integrated Trajectory Prediction.** We jointly predict agent multimodal intentions and trajectories in a unified framework, eliminating redundant modules and enhancing information flow.

2) **Context-Aware Pruning via Dual Filters.** We introduce complementary agent and map filters that leverage predicted behavioral interaction probabilities and vectorized occupancy to retain only influential vehicles and relevant map elements.

3) **Automatic Intention Label Generation.** We propose an automatic labeling strategy to annotate agent-level intentions in large-scale datasets (e.g., Waymo, Argoverse). This strategy enables more convenient behavior prediction without manual effort.

4) **State-of-the-Art Performance and Real-World Validation.** IMPACT achieves SOTA results on mainstream public benchmarks and demonstrates robust real-world performance through deployment on an autonomous vehicle (a video demo in supplementary material).

II. Related Works

A. Motion Prediction

Current methods [1]–[4] tend to emphasize the precise generation of multimodal trajectory predictions at the expense of behavioral intention prediction [13] (BIP). This narrow focus often overlooks the latent decision-making processes that drive observable maneuvers, resulting in models that struggle to interpret complex social interactions or anticipate nuanced changes in driving behavior. In real-world scenarios, the inability to infer intentions such as overtake or yield decisions can lead to suboptimal planning and reduced overall robustness. The most recent work, BeTOP [11], attempts to explicitly model behavioral interactions via braid theory by splitting agents into "interactive" and "non-interactive" based on whether their future trajectories form an intertwined braid. However, in most scenarios, the generated braids appear unrealistic because the approach only considers lateral and temporal dimensions. Moreover, relying on a binary behavioral label oversimplifies complex traffic scenarios, where interactions can be far more nuanced.

Our IMPACT framework addresses these limitations by jointly predicting multimodal, multi-class behavioral intentions and future trajectories. To better supervise the behavioral intention predictor, we propose an auto-labeling algorithm that generates reasonable intention labels, striking a balance between explainability and behavioral completeness.

B. Context-aware Pruning

During the decoder stage, current works [3], [7]–[10], [13]–[17] often relies on attention mechanisms to query agent-map information. However, using global attention over all entities incurs $O(n^2)$ complexity and can introduce causal confusion, in contrast to human drivers, who primarily focus on goal-critical paths and potentially interactive vehicles [13]. Some approaches address this via prior rule-based pruning (e.g., SEPT [7]), which may fail for corner case. Meanwhile, MTR-series [3], [8]–[10], [14]–[16] methods and R-pred [6] use last layer predictions as priors to dynamically collect nearby polylines, yet suffers from error propagation. BeTOP [11], which relies on braid theory to select interactive vehicles, can yield an unreasonably chosen set of vehicles in practice. To tackle these issues, our IMPACT framework introduces a symmetric dual context-filtering approach that leverage predicted behavioral interaction probabilities and vectorized occupancy to retain only influential vehicles and relevant map elements.

III. Methodologies

A. Problem Formulation

The historical trajectories of N_a traffic participants are denoted as $\mathcal{A} = \{a_1, a_2, \ldots, a_{N_a}\}$. The corresponding map is equally partitioned into N_l polylines $\mathcal{L} = \{l_1, l_2, \ldots, l_{N_l}\}$. The predictor will anticipate K different modality future trajectories $\mathcal{Y} = \{y_1, y_2, \ldots, y_{N_a}\}$ over the future T_f timesteps, where $y_i = \{y_i^1, y_i^2, \ldots, y_i^K\} \in \mathbb{R}^{K \times T_f \times 2}$. The confidence score for y_i are denoted as $s_i = \{s_i^1, s_i^2, \ldots, s_i^K\}$. Then for target agent a_i, existing motion prediction task aims to estimate the distribution:

$$P(y_i \mid \mathcal{L}, \mathcal{A}) = \sum_{k=1}^{K} s_i^k P(y_i^k \mid \mathcal{L}, \mathcal{A}) \tag{1}$$

To better capture inter-agent interactions and crucial map segments, we additional predict two sets of modality-dependent priors: behavioral intentions $\mathcal{H}^k = \{h_1^k, \ldots, h_{N_a}^k\}$, where $h_i^k \in \mathbb{R}^4$ encodes the probability distribution of behavioral intentions (overtaking, yielding, ignored, nearby) for agent a_i toward target agent under mode k and vectorized occupancy $\mathcal{O}^k = \{o_1^k, \ldots, o_{N_\ell}^k\}$, where $o_j^k \in [0, 1]$ indicates

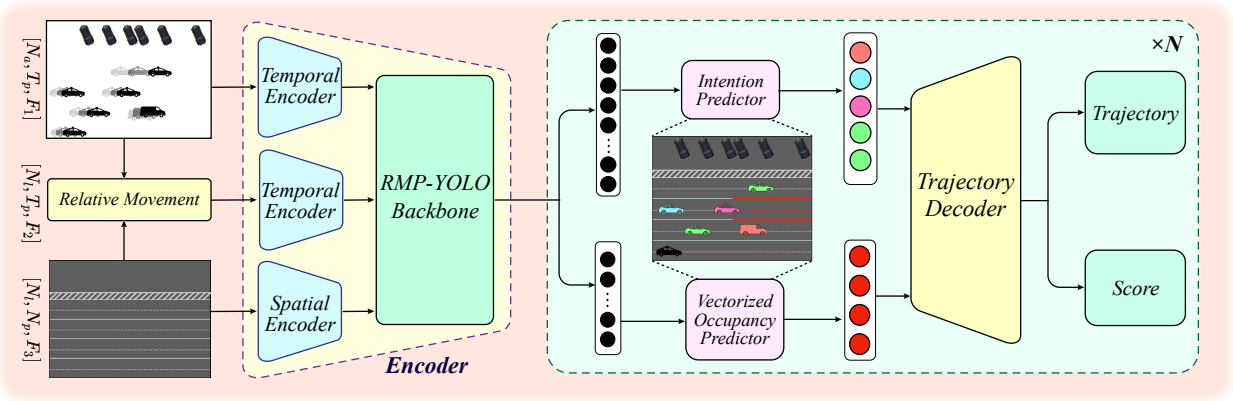

Fig. 2. An overview of framework of IMPACT. Both the Intention Predictor and the Vectorized Occupancy Predictor share the same context encoder with the Trajectory Decoder, leveraging their outputs to prune irrelevant agents and map polylines. This selective mechanism ensures that only the most critical context is fed into the decoder for final trajectory prediction.

the probability that polyline l_j is relevant or "occupied" by the agent's future path under mode k. Given $\{\mathcal{H}^k\}$ and $\{\mathcal{O}^k\}$, we perform a top-m selection of agents and top-n selection of map polylines for each modality k:

$$\begin{aligned}\mathcal{A}_k^{\text{sel}} &= \Big\{a_j :\ j \in \text{argtop}_m\big[\psi(h_j^k)\big]_{j=1}^{N_a}\Big\},\\ \mathcal{L}_k^{\text{sel}} &= \Big\{l_j :\ j \in \text{argtop}_n\big[\varphi(o_j^k)\big]_{j=1}^{N_\ell}\Big\}.\end{aligned} \tag{2}$$

where $\psi(h_j^k)$ and $\varphi(\mathbf{o}_j^k)$ is a scalar score derived from the intention vector $\mathbf{h}_j^k$ and occupancy vector $\mathbf{o}_j^k$ by the utility function $\psi(\cdot)$ and $\varphi(\cdot)$. These subsets ensure that each modality k focuses only on the most critical agents and polylines.

To jointly capture both the behavioral intentions and the map occupancy, we consider the extended distribution:

$$\begin{aligned}&P\Big(y_i,\ \mathcal{H},\ \mathcal{O}\ \Big|\ \mathcal{L}, \mathcal{A}\Big) \approx\\ &\sum_{k=1}^{K} s_i^k P\big(\mathcal{H}^k \mid \mathcal{L}, \mathcal{A}\big)\, P\big(\mathcal{O}^k \mid \mathcal{L}, \mathcal{A}\big)\ P\Big(y_i^k \mid \mathcal{A}_k^{\text{sel}}, \mathcal{L}_k^{\text{sel}}\Big).\end{aligned} \tag{3}$$

Where $P\big(\mathcal{H}^k \mid \mathcal{L}, \mathcal{A}\big)$ yields the intention vectors for each agent under mode k, $P\big(\mathcal{O}^k \mid \mathcal{L}, \mathcal{A}\big)$ produces occupancy scores for each polyline under mode k, $\mathcal{A}_k^{\text{sel}}$ and $\mathcal{L}_k^{\text{sel}}$ are the selected agents/polylines for each modality, and the final trajectory y_i^k is decoded by attending only to these subsets. This operation reduces cross-attention complexity from $O(KN_a{+}KN_\ell)$ to $O(Km{+}Kn)$ while maintaining accuracy.

B. Input Representation

In our method, we apply agent-centric normalization. To predict a target agent, the input to the predictor comprises: $\mathcal{A} = \{a_1, a_2, \ldots, a_{N_a}\} \in \mathbb{R}^{N_a \times T_p \times F_1}$, 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.

Based on the approach of RMP-YOLO [10], 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$((\Delta x,\ \Delta y,\ \cos\Delta\theta,\ \sin\Delta\theta))$.

C. Network Structure

The vectorized input is first processed by a spatio-temporal encoder to generate informative instance-level tokens. Subsequently, symmetric dual filters are applied for context-aware pruning, ensuring that only influential agent tokens and map polyline tokens that are likely to be occupied in the future are forwarded to the trajectory decoder. These selected tokens are then forwarded to the trajectory decoder, enabling efficient and focused multimodal motion prediction.

1) Spatial-temporal Encoding: We employ the encoder from RMP-YOLO [10] as our backbone network. Specifically, two temporal encoders and one spatial encoder are first utilized to tokenize different input modalities (see Fig.2). Local attention mechanisms are then applied to propagate spatial-temporal information among the resulting tokens (More details can be found in our supplementary video). The output of this stage consists of compact yet informative agent tokens $\mathcal{A}_1 \in \mathbb{R}^{N_a \times D}$ and map tokens $\mathcal{L}_1 \in \mathbb{R}^{N_l \times D}$:

$$\mathcal{A}_1, \mathcal{L}_1 = \text{Encoder}(\mathcal{A}, \mathcal{R}, \mathcal{L}) \tag{4}$$

Before diving into decoder part (see Fig. 3), 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. K denotes the number of different futures.

2) 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 $\mathcal{A}_1 \in \mathbb{R}^{N_a \times D}$ 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:

$$\begin{aligned}\hat{\text{H}}^i &= \text{Softmax}\Big(\text{MLP}\big(I^i\big)\Big) \in \mathbb{R}^{K \times N_a \times 4},\\ I^i &= \text{MLP}(\text{MLP}(\mathcal{A}_1 \oplus Q^i) + I^{i-1}).\end{aligned} \tag{5}$$

Each vector element h represents a probability distribution over four intention categories: *yielding*, *overtaking*, *ignored*,

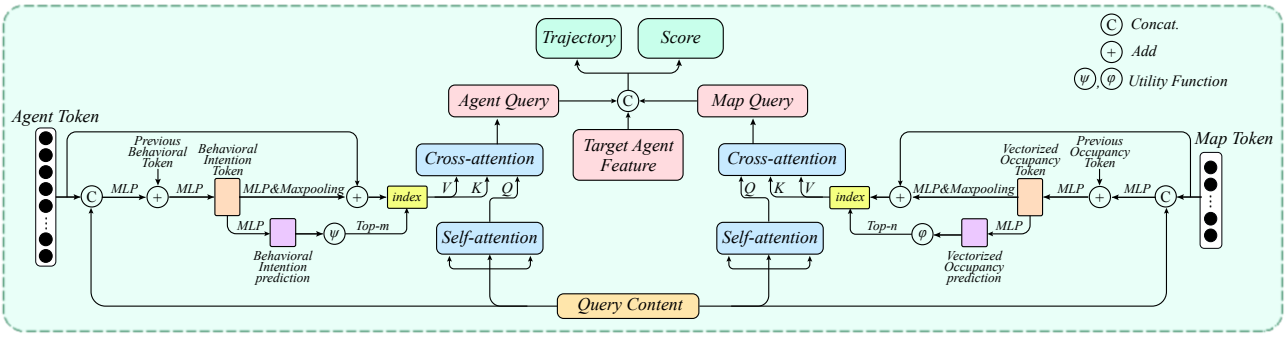

Fig. 3. An overview of our decoder framework, featuring context-aware pruning via symmetric dual filters.

Algorithm 1 GT behavioral Intention Label Generation

Input:
Agents' future trajectories $Y^* \in \mathbb{R}^{N_a \times T_f \times 2}$;
Target agent's future trajectory $y \in \mathbb{R}^{T_f \times 2}$;
Output:
behavioral Intention labels $H \in \mathbb{R}^{N_a \times 4}$;

Radius Screening:
for $i = 1 \dots N_a$ **do**
  Compute $d_i(t) = \|Y^*_{i,t} - y_t\|_2, \ \forall\, t = 1 \dots T_f$.
  if $\min_t\, d_i(t) > \tau_{\text{ignore}}$ **then**
    Label i as `ignored` and exclude from further checks.
  else
    Add i to the candidate set $\mathcal{C}$.
  end if
end for
Interaction Detection:
for $i \in \mathcal{C}$ **do**
  Define geometry $S_i(t)$ for agent i, and $S_{\text{target}}(t)$ for the target agent.
  If $\Big(\forall\, t : \ S_i(t) \cap S_{\text{target}}(t) = \varnothing\Big) \wedge \big(\min_t d_i(t) \le \tau_{\text{ignore}}\big)$: label agent i as `nearby`.
  Else if $\exists\, t : \ S_i(t) \cap S_{\text{target}}(t) \neq \varnothing$: compare the time steps of minimal distances to assign either `overtaking` or `yielding`.
end for
Label Assignment:
Convert each final label {`ignored`, `nearby`, `overtaking`, `yielding`} into a one-hot vector $H_i \in \mathbb{R}^4$.
return H

Algorithm 2 Vectorized GT Occupancy Label Generation

Input:
Map polyline points $L \in \mathbb{R}^{N_l \times N_p \times 2}$;
Target agent's future trajectory $y \in \mathbb{R}^{T_f \times 2}$;
Distance threshold α;
Output:
Occupancy labels $O \in \{0, 1\}^{N_l \times 1}$;

Broadcast L to $[\,N_l,\, N_p,\, 1,\, 2\,]$ and y to $[\,1,\, 1,\, T_f,\, 2\,]$.
Compute the distance tensor $d \in \mathbb{R}^{N_l \times N_p \times T_f}$.
$O = (\min(d, \text{axis} = 1) \le \alpha\ ).\text{any}(\text{axis} = -1)$.
return $O \in \mathbb{R}^{N_l \times 1}$

and *nearby*. To make the decoding process 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 $\mathcal{A}_2 \in \mathbb{R}^{m \times D}$, 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 (see Algorithm 1).

3) 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:

$$\begin{aligned} \hat{\mathrm{O}}^i &= \text{Sigmoid}\Big(\text{MLP}\big(C^i\big)\Big) \ \in \ \mathbb{R}^{K \times N_l \times 1}, \\ C^i &= \text{MLP}(\text{MLP}(\mathcal{L}_1 \oplus Q^i) + C^{i-1}) \end{aligned} \tag{6}$$

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 $\mathcal{L}_2 \in \mathbb{R}^{n \times D}$, 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 (see Algorithm 2).

4) Trajectory Decoder: We adopt a multi-layer MTR-style [8] 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 $\mathcal{A}_2$ and map tokens $\mathcal{L}_2$. 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.

D. Training Loss

Our overall training objective comprises four components:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{Int}} + \lambda_2 \mathcal{L}_{\text{Occ}} + \lambda_3 \mathcal{L}_{\text{Traj}} + \lambda_4 \mathcal{L}_{\text{Score}}, \quad (7)$$

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 $\lambda_4 = 1$.

IV. Experiments

A. Experimental Setup

1) Datasets and Evaluation Metrics: Our experiments are conducted on one of the most challenging prediction datasets, Waymo Open Motion Dataset (WOMD) [18]. This large-scale dataset comprises 486,995 training clips, 44,097 validation clips, and 44920 testing clips. Each clip contains 10 timesteps of historical agent states, 1 current timestep, and 80 future timesteps at a sampling frequency of 10 Hz, along with HD map information. We evaluate our method on both core WOMD tasks: the marginal motion and interactive motion prediction tasks. Following the standard evaluation protocol, we adopt metrics including Soft mAP, mAP, minADE, minFDE, Miss Rate, and Overlap Rate, and Soft mAP is the main metric.

2) Implementation Details.: We employ AdamW optimizer [19] 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.

B. Leaderboard Performance

1) Marginal Prediction Performance: Table I shows that our method achieves state-of-the-art (SOTA) performance on the Waymo motion prediction benchmark, outperforming all existing LiDAR-free approaches in the primary metric, Soft mAP, as well as in minADE and Overlap Rate. In the remaining metrics (mAP, minFDE, and Miss Rate), our method ranks second. Here, 'Ensemble' denotes the model-ensemble technique described in [8] for performance boosting, while 'Single' indicates evaluation with a single model.

2) Joint Prediction Performance: As presented in Table II, even without any model-ensemble techniques, our single model achieves the best performance on the Waymo joint prediction leaderboard, surpassing the previous SOTA method [11] by 10.2% in both Soft mAP and mAP. Furthermore, our model attains the second-lowest Overlap Rate and the third-lowest Miss Rate. These substantial improvements underscore the effectiveness of the IMPACT framework.

C. Behavioral Intention Prediction Performance

TABLE III
Multimodal Behavioral Intention Prediction Performance.

Class	Precision	Top-1 Recall	F1-Score	Top-6 Accuracy	GT Data Ratio (%)
Ignored	0.99	0.97	0.98	0.99	89.12
Nearby	0.71	0.89	0.79	0.97	3.93
Overtaking	0.82	0.89	0.85	0.97	2.59
Yielding	0.86	0.96	0.90	0.97	4.32
All	0.97	0.97	0.97	0.987	100.00

Table III presents the results of multimodal behavioral intention prediction across four designated categories. Overall, the model demonstrates strong performance, with a F1-Score of 0.97, and a Top-6 accuracy of 0.987. In the dominant *Ignored* category, which comprises 89.12% of the dataset, our method attains particularly high accuracy (F1 = 0.98). Despite being underrepresented, the *Nearby*, *Overtaking*, and *Yielding* classes achieve F1-Scores between 0.79 and 0.90, showcasing the model's robustness in handling less frequent behaviors. These results underscore the model's effectiveness in accurately identifying and predicting diverse driver intentions, offering valuable insights for interpretable prediction models and informed downstream decision-making. Per-mode visualization results are provided in Figure 4.

D. Vectorized Occupancy Prediction Performance

TABLE IV
Binary Classification Performance for top-1 mode Occupancy Prediction.

Class	Precision	Recall	F1-Score	GT Data Ratio (%)
Occupied	0.933	0.775	0.847	3.96
Unoccupied	0.991	0.998	0.994	96.04
All		0.989		100.00

Table IV presents precision, recall, and F1-score for both classes, along with their data ratios in the validation set. The *occupied* class constitutes only 3.96% of samples, highlighting a strong imbalance. Despite this, the model achieves a high overall accuracy of 0.989. For the *occupied* class, precision (0.933) and recall (0.775) yield an F1-score of 0.847, indicating rare false positives—crucial for pruning irrelevant map polylines in trajectory prediction. The dominant *unoccupied* class achieves near-perfect precision (0.991), recall (0.998), and F1-score (0.994). These results demonstrate that our vectorized occupancy prediction effectively identifies critical map segments while minimizing false classification.

TABLE I

PREDICTION ON THE TEST LEADERBOARD OF THE MOTION PREDICTION TRACK OF THE WAYMO OPEN DATASET CHALLENGE. THE FIRST PLACE IS DENOTED BY **BOLD**, THE SECOND PLACE BY UNDERLINE, AND THE THIRD PLACE BY ∗ASTERISK.

Method	Soft mAP ↑	mAP ↑	minADE ↓	minFDE ↓	Miss Rate ↓	Overlap Rate ↓
RMP-YOLO(Ensemble) [10]	0.4737	0.4531	0.5564	**1.1188**	**0.1084**	0.1259*
ModeSeq(Ensemble) [2]	0.4737	**0.4665**	0.5680	1.1766	0.1204	0.1275
BeTop [11]	0.4698	0.4587	0.5716	1.1668	0.1183	0.1272
MGTR [15]	0.4599	0.4505	0.5918	1.2135	0.1298	0.1275
EDA [9]	0.4596	0.4487	0.5718	1.1702	0.1169	0.1266
MTR++ [3]	0.4410	0.4329	0.5906	1.1939	0.1298	0.1281
MTR [8]	0.4403	0.4249	0.5964	1.2039	0.1312	0.1274
HDGT [20]	0.3709	0.3577	0.5933	1.2055	0.1511	0.1557
DenseTNT [21]	-	0.3281	1.0387	1.5514	0.1573	0.1779
SceneTransformer [22]	-	0.2788	0.6117	1.2116	0.1564	0.1473
Ours (Ensemble)	**0.4801**	0.4598*	**0.5563**	1.1295	0.1087	0.1258
Ours (Single)	0.4721	0.4609	0.5641*	1.1540*	0.1143*	**0.1255**

TABLE II

JOINT PREDICTION ON THE TEST LEADERBOARD OF THE INTERACTION PREDICTION TRACK OF THE WAYMO OPEN DATASET CHALLENGE. THE FIRST PLACE IS DENOTED BY **BOLD**, THE SECOND PLACE BY UNDERLINE, AND THE THIRD PLACE BY ∗ASTERISK.

Method	Soft mAP ↑	mAP ↑	minADE ↓	minFDE ↓	Miss Rate ↓	Overlap Rate ↓
BeTop-ens [11]	0.2573	0.2511	0.9779	2.2805	0.4376	0.1688*
BeTop [11]	0.2466*	0.2412*	0.9744	2.2744	0.4355	0.1696
MTR++ [3]	0.2368	0.2326	0.8975	1.9509	**0.4143**	**0.1665**
MTR [8]	0.2078	0.2037	0.9181*	2.0633	0.4411	0.1717
MotionDiffuser [23]	0.2047	0.1952	**0.8642**	**1.9482**	0.4300	0.2004
GameFormer [24]	0.1982	0.1923	0.9721	2.2146	0.4933	0.2022
DenseTNT [21]	-	0.1647	1.1417	2.4904	0.5350	0.2309
M2I [25]	-	0.1239	1.3506	2.8325	0.5538	0.2757
SceneTransformer [22]	-	0.1192	0.9774	2.1892	0.4942	0.2067
Ours (Single)	**0.2718**	**0.2659**	0.9738	2.2734	0.4316*	0.1684

E. Ablation Study

1) Different Behavior Modeling and Map Interaction Mechanisms: Table V presents an ablation study evaluating various agent-pruning and map-pruning strategies on the validation set. The first row (All + Dynamic) queries all agents for the decoder while employing dynamic map selection as in MTR [8], serving as a baseline. Introducing agent selection via Braid Theory [11] (Row 2) enhances Soft mAP to 0.4602. Replacing Braid Theory with our two-class filtering—where Yielding, Overtaking, and Nearby behaviors are treated as interactive while Ignore is considered non-interactive (Row 3)—further refines minADE to 0.5693 while maintaining a comparable Miss Rate. Notably, adopting our four-class YOIN approach (Row 4) elevates Soft mAP to 0.4728, demonstrating its ability to better capture nuanced interactions. Lastly, incorporating vectorized occupancy into YOIN (Row 5) further improves performance (Soft mAP = 0.4758), highlighting the benefit of explicit map filtering in isolating relevant polylines and refining trajectory predictions. These results validate the effectiveness of our integrated behavior modeling and map selection framework.

2) Computational Efficiency and Parameter Analysis: We compare various agent-pruning and map-pruning strategies in Table V using the same encoder architecture. The first row (All + w/Dynamic) corresponds to a baseline that queries all agents and dynamically collect polylines, resulting in a model with 69.469 M parameters and a reported inference time of 14.67 ms under a simplified setting. By contrast, our final approach (the last row in combines YOIN-based agent pruning with vectorized occupancy-based pruning. This design lowers the parameter count to 48.944 M while reducing the average inference time of BeTOP [11] from 23.92 to 18.27 ms per scenario. The key to these efficiency gains lies in pruning irrelevant scene information—both agents and polylines—based on their semantic contributions, which in turn eliminates redundant computations.

3) Number of Selected Tokens: As shown in Figure 5, selecting 24 agents and 192 polylines achieves the optimal balance between interaction diversity and trajectory accuracy. On average, the model input contains 43.85 agents and 749.52 map polylines, meaning this selection reduces scene complexity by 45.2% and 74%, respectively, while preserving essential contextual cues. Beyond this threshold, redundant elements introduce noise, degrading generalization and increasing prediction uncertainty. This highlights the importance of selective context pruning to enhance trajectory forecasting performance.

4) Generalization Ability: To validate the cross-dataset and cross-paradigm generalization of our proposed symmetric dual filter, we conduct transfer experiments between the Waymo Open Dataset and Argoverse (AV1/AV2) using two distinct baselines: SIMPL (query-centric) and HiVT (agent-centric). As shown in Table VI, our approach achieves up to 7.1% (SIMPL on AV2) and 6.4% (HiVT on AV1) improvements in bFDE_6. These results demonstrate our method's versatility

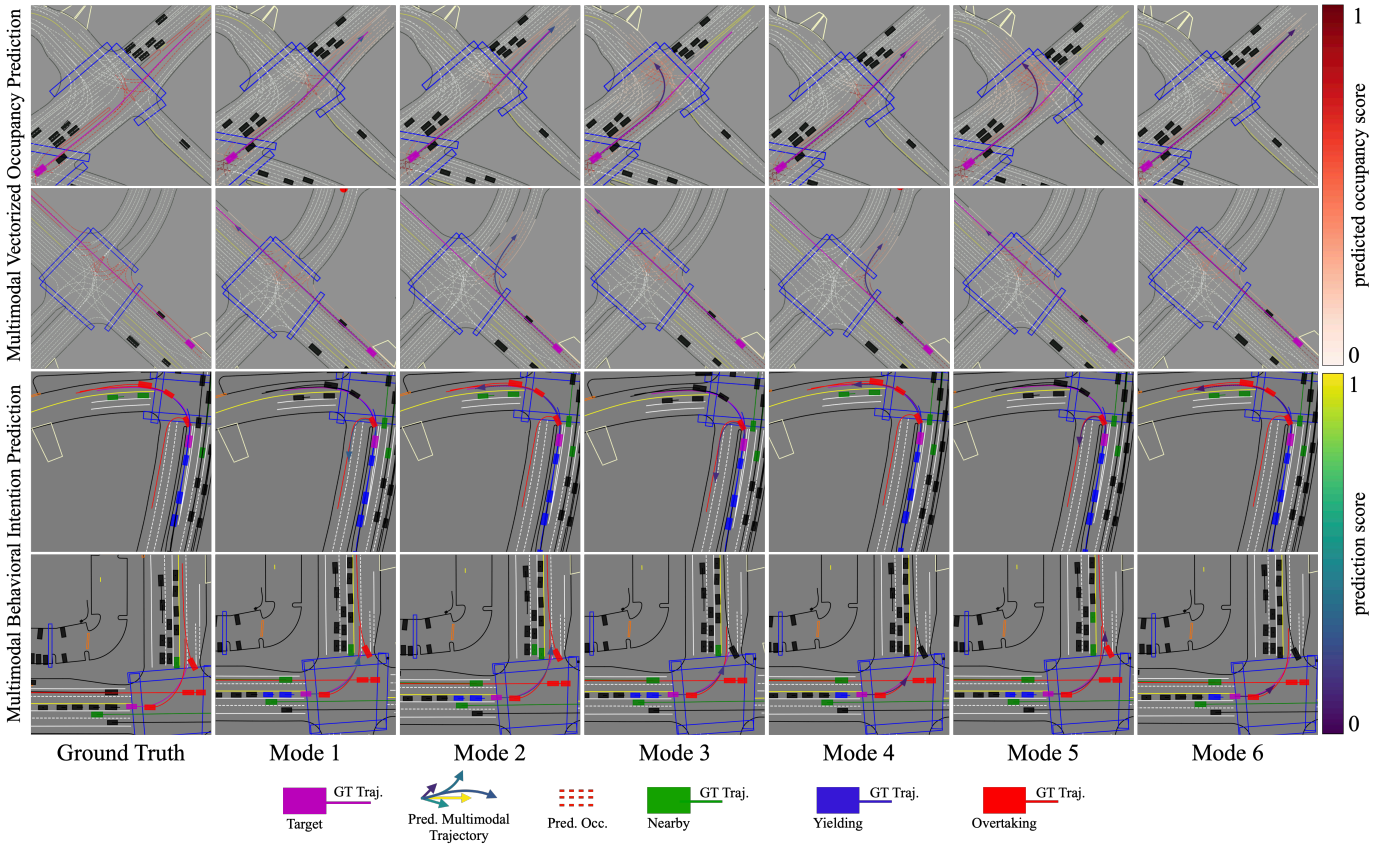

Fig. 4. Visualization of Predicted Multimodal Occupancy and Intention Labels. In the top two rows, black agents represent other agents, while in the bottom two rows, they indicate ignored agents. Ground-truth trajectories are included for validation of predicted behaviors.

TABLE V
ABLATION STUDY OF DIFFERENT AGENT-PRUNING AND MAP-PRUNING STRATEGIES ON THE VALIDATION SET, USING THE SAME PROPOSED ENCODER.

Agent			Map			Metrics					
All	w/Braid Theory [11]	w/Inter.	w/YOIN	w/Dynamic [8]	w/Vect. Occ.	Soft mAP	mAP	minADE	Miss Rate	Inference Time (ms/scenario)	Params (M)
✓				✓		0.4582	0.4414	0.5718	0.1203	14.67	69.469
	✓			✓		0.4602	0.4507	0.5745	0.1187	23.92	48.938
		✓		✓		0.4663	0.4549	0.5693	0.1187	23.92	48.938
			✓	✓		0.4728	0.4613	0.5680	0.1163	23.92	48.941
			✓		✓	0.4758	0.4646	0.5652	0.1173	18.27	48.944

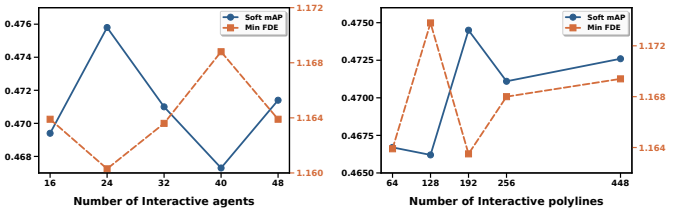

Fig. 5. Results of varying numbers of selected agents and polylines for modeling on validation dataset.

TABLE VI
PERFORMANCE GAINS FROM INTEGRATING PROPOSED MODULES WITH EXISTING APPROACHES (ARGOVERSE1&2). BFDE6 IS THE MAIN METRIC.

Method	bFDE_6	mADE_6	mFDE_6	MR
SIMPL (AV2)	2.069	0.777	1.452	0.196
SIMPL+Ours (AV2)	1.921 (−7.1%)	0.743	1.387	0.178
HiVT (AV1)	1.662	0.661	0.969	0.092
HiVT+Ours (AV1)	1.556 (−6.4%)	0.599	0.932	0.087

as a general enhancement framework, independent of specific network architectures.

V. CONCLUSION

In this paper, we present IMPACT, a novel and unified module that advances multimodal motion prediction through explicit modeling of behavioral intentions and dynamic context optimization. The joint intention and motion modeling module eliminates redundancy and enables seamless information flow between behavioral semantic and motion dynamics, the experiments on both marginal and joint motion prediction challenges of large-scale WOMD show that our approach achieves state-of-the-art performance. The adaptive pruning decoder leverages intention and occupancy prediction priors to reduce computational complexity while preserving essential interaction cues. The automated labeling framework generates intention annotations across mainstream datasets and shows great scalable performance.

Limitations and Future work. Our pruning mechanism prioritizes instantaneous interactions, potentially overlooking

evolving multi-agent gaming dynamics. Extending the framework with recursive reasoning could enhance long-horizon interaction modeling. We plan to incorporate this idea into both the latest end-to-end and traditional planning frameworks, particularly to address the challenge of determining which agent to predict and which polyline to focus on.

REFERENCES

[1] M. Wang, X. Ren, R. Jin, M. Li, X. Zhang, C. Yu, M. Wang, and W. Yang, "Futurenet-lof: Joint trajectory prediction and lane occupancy field prediction with future context encoding," 2024. [Online]. Available: https://arxiv.org/abs/2406.14422

[2] Z. Zhou, H. Zhou, H. Hu, Z. Wen, J. Wang, Y.-H. Li, and Y.-K. Huang, "Modeseq: Taming sparse multimodal motion prediction with sequential mode modeling," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, June 2025, pp. 1612–1621.

[3] S. Shi, L. Jiang, D. Dai, and B. Schiele, "Mtr++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3955–3971, 2024.

[4] 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.

[5] Y. Zhou, H. Shao, L. Wang, S. L. Waslander, H. Li, and Y. Liu, "Smartrefine: A scenario-adaptive refinement framework for efficient motion prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 15 281–15 290.

[6] S. Choi, J. Kim, J. Yun, and J. W. Choi, "R-pred: Two-stage motion prediction via tube-query attention-based trajectory refinement," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 8525–8535.

[7] Z. Lan, Y. Jiang, Y. Mu, C. Chen, and S. E. Li, "SEPT: Towards efficient scene representation learning for motion prediction," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=efeBC1sQj9

[8] 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*, vol. 35. Curran Associates, Inc., 2022, pp. 6531–6543.

[9] L. Lin, X. Lin, T. Lin, L. Huang, R. Xiong, and Y. Wang, "Eda: Evolving and distinct anchors for multimodal motion prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 4, 2024, pp. 3432–3440.

[10] 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

[11] H. Liu, L. Chen, Y. Qiao, C. Lv, and H. Li, "Reasoning multi-agent behavioral topology for interactive autonomous driving," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. [Online]. Available: https://openreview.net/forum?id=FSgwgQXTxo

[12] C. Shi, S. Shi, and L. Jiang, "Mtr v3: 1st place solution for 2024 waymo open dataset challenge - motion prediction," The Chinese University of Hong Kong (Shenzhen) and DiDi Global, Technical Report, 2024. [Online]. Available: https://storage.googleapis.com/waymo-uploads/files/research/2024%20Technical%20Reports/2024%20WOD%20Motion%20Prediction%20Challenge%20-%201st%20Place%20-%20MTR%20v3.pdf

[13] J. Fang, F. Wang, J. Xue, and T.-S. Chua, "Behavioral intention prediction in driving scenes: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 8, pp. 8334–8355, 2024.

[14] J. Sun, C. Yuan, S. Sun, S. Wang, Y. Han, S. Ma, Z. Huang, A. Wong, K. P. Tee, and M. H. Ang, "Controlmtr: Control-guided motion transformer with scene-compliant intention points for feasible motion prediction," in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, 2024, pp. 1507–1514.

[15] Y. Gan, H. Xiao, Y. Zhao, E. Zhang, Z. Huang, X. Ye, and L. Ge, "Multi-granular transformer for motion prediction with lidar," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 15 092–15 098.

[16] X. Zheng, L. Wu, Z. Yan, Y. Tang, H. Zhao, C. Zhong, B. Chen, and J. Gong, "Large language models powered context-aware motion prediction in autonomous driving," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 980–985.

[17] J. Li, T. Shen, Z. Gu, J. Sun, C. Yuan, Y. Han, S. Sun, and M. H. Ang, "Adm: Accelerated diffusion model via estimated priors for robust motion prediction under uncertainties," in *2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC)*, 2024, pp. 2221–2227.

[18] Waymo, "Waymo open dataset: Motion prediction challenge," accessed: 2024-08-31. [Online]. Available: https://waymo.com/open/challenges

[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[20] X. Jia, P. Wu, L. Chen, Y. Liu, H. Li, and J. Yan, "Hdgt: Heterogeneous driving graph transformer for multi-agent trajectory prediction via scene encoding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13 860–13 875, 2023.

[21] J. Gu, C. Sun, and H. Zhao, "Densetnt: End-to-end trajectory prediction from dense goal sets," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 15 283–15 292. [Online]. Available: https://doi.org/10.1109/ICCV48922.2021.01502

[22] J. Ngiam, V. Vasudevan, B. Caine, Z. Zhang, H. L. Chiang, J. Ling, R. Roelofs, A. Bewley, C. Liu, A. Venugopal, D. J. Weiss, B. Sapp, Z. Chen, and J. Shlens, "Scene transformer: A unified architecture for predicting future trajectories of multiple agents," in *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. [Online]. Available: https://openreview.net/forum?id=Wm3EA5OlHsG

[23] C. ". Jiang, A. Cornman, C. Park, B. Sapp, Y. Zhou, and D. Anguelov, "Motiondiffuser: Controllable multi-agent motion prediction using diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 9644–9653.

[24] Z. Huang, H. Liu, and C. Lv, "Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 3903–3913.

[25] Q. Sun, X. Huang, J. Gu, B. C. Williams, and H. Zhao, "M2i: From factored marginal trajectory prediction to interactive prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 6543–6552.