# MTR++: 1st Place Solution for 2023 Waymo Open Dataset Challenge - Motion Prediction

Shaoshuai Shi* Li Jiang* Dengxin Dai Bernt Schiele

Max Planck Institute for Informatics, Saarland Informatics Campus

{sshi, lijiang, ddai, schiele}@mpi-inf.mpg.de

## Abstract

*This report presents our solution that won the 1st place in the Waymo Open Dataset (WOD) Motion Prediction Challenge 2023. Building upon the previous year's champion approach, MTR, we propose an enhanced version called MTR++. This improved framework enables simultaneous prediction of future trajectories for multiple agents by incorporating two novel strategies: symmetric scene context modeling and mutually-guided intention querying. To further enhance our performance on the leaderboard, we adopt the same model ensemble strategy as MTR. This strategy involves merging predictions from multiple models using non-maximum suppression. Finally, our approach achieves the $1^{st}$ place on the motion prediction leaderboard of the WOD Challenges 2023, highlighting its effectiveness.*

## 1. Introduction

Motion prediction is a pivotal undertaking in the realm of contemporary autonomous driving systems, enabling robotic vehicles to comprehend complex driving scenarios and make judicious decisions [1, 7, 3, 9, 4, 2]. However, this task presents great challenges due to the inherently multimodal behaviors exhibited by traffic participants and the complex nature of the surrounding environmental contexts.

To address these challenges, we base our solution on the state-of-the-art approach MTR [5], which emerged as the champion in last year's challenge. While MTR achieves strong performance in motion prediction, it focuses on generating future trajectories for a single focal agent. In contrast, we propose an advanced framework called MTR++, which enables the generation of multimodal future trajectories for multiple agents simultaneously.

To achieve this, we introduce a novel symmetric scene context modeling strategy in MTR++. Instead of encoding scene context features based on the global coordinate system of the focal agent as in MTR, we encode the features for each token individually in their local coordinate system using a query-centric self-attention module. These symmetrically encoded token features can be utilized for the motion prediction of any agent. Additionally, we employ a joint motion decoder in MTR++ to generate future trajectories for multiple focal agents simultaneously. Each focal agent has its own intention queries, which are used to generate its corresponding multimodal future trajectories. To enhance prediction accuracy and enable agents to interact and influence each other's behavior, we propose the mutually-guided intention querying strategy based on the motion decoder of MTR. This involves adding a self-attention layer to facilitate information propagation among the intention queries from different focal agents before feeding them into the transformer decoder layer. Thanks to these two novel improvements, the experiments show that MTR++ not only achieves better performance but also enables the simultaneous motion prediction of multiple agents.

## 2. Method

The overall architecture of our approach is based on the MTR framework [5], and we introduce two novel improvements: the symmetric scene context modeling in the context encoder and the mutually-guided intention querying in the motion decoder. These new contributions are illustrated in the following paragraphs.

### 2.1. Symmetric Scene Context Modeling

Unlike most existing methods that focus on a specific agent and center the scene around them [10, 1, 8, 5], our approach symmetrically encodes the entire scene for each agent. This enables the encoded scene context features to be directly utilized for predicting the motion of any agent by attaching a motion decoder network.

Specifically, to encode the input context features, we employ the same vectorized representation as in MTR. However, instead of normalizing all inputs to a global coordinate

This is a non-peer-reviewed technical report for the motion prediction track of 2022 Waymo Open Dataset Challenge.

system centered on a focal agent, we encode the feature of each polyline in a polyline-centric local coordinate system. The local coordinate system for each agent's corresponding polyline is determined based on the agent's latest position and moving direction. For map polylines, the local coordinate system is determined based on the geometry center and tangent direction of each polyline. This polyline encoding process generates the agent features $A \in \mathbb{R}^{N_a \times D}$ and the map features $M \in \mathbb{R}^{N_m \times D}$, where $N_a$ is the number of agents, $N_m$ is the number of map polylines, and $D$ is the feature dimension.

To model the relationships between tokens, MTR uses a native self-attention module that relies on a global coordinate system centered on the focal agent. In contrast, we propose the query-centric self-attention module to model the tokens' relationships in a symmetric manner without relying on any global coordinate system. This module performs the attention mechanism separately for each query token, allowing us to explore the relationship between a query token and other tokens in its specific local coordinate system. For example, when considering the $i$-th token as the query, we convert the coordinates and directions of all tokens into the local coordinate system of the query token, as follows:

$$\begin{aligned} F'^{(\mathrm{l})}_{\mathrm{AM}}[i] = \mathrm{MHSA}\big(&\mathrm{Q}\colon [F^{(\mathrm{l})}_{\mathrm{AM}}[i], \mathrm{PE}(R_{\mathrm{AM}[i,i]})], \\ &\mathrm{K}\colon \big\{[F^{(\mathrm{l})}_{\mathrm{AM}}[j], \mathrm{PE}(R_{\mathrm{AM}[i,j]})]\big\}_{j\in\Omega(i)}, \\ &\mathrm{V}\colon \big\{F^{(\mathrm{l})}_{\mathrm{AM}}[j] + \mathrm{PE}(R_{\mathrm{AM}[i,j]})\big\}_{j\in\Omega(i)}\big), \end{aligned} \tag{1}$$

where $i \in \{1, \ldots, N_a + N_m\}$, and $j \in \Omega(i)$ indicating the index of its neighboring tokens. $F^{(\mathrm{l})}_{\mathrm{AM}[i]}$ is the input features of the $i$-th token. $R_{\mathrm{AM}[i,j]}$ indicate the $j$-th token's relative position and direction in the local coordinate system of the $i$-th query token, and $\mathrm{PE}(\cdot)$ indicates the sinusoidal positional encoding.

By employing the proposed query-centric self-attention module, we achieve symmetric encoding of scene context features for each input token. This enables the encoded features to be effectively utilized for predicting the motion of any input agent.

### 2.2. Mutually-Guided Intention Querying

With the symmetrically-encoded scene context features, MTR++ leverages MTR's motion decoder to simultaneously predict future trajectories for multiple agents. To further enhance the performance of multi-agent motion prediction, we propose a mutually-guided intention querying strategy that enables agents to interact and influence each other's behavior. This is achieved by introducing a query-centric self-attention layer among the intention queries from different focal agents.

More specifically, we represent the intention queries in MTR++ as $E^{(\mathrm{m})}_{\mathrm{I}} \in \mathbb{R}^{N_o \times \mathcal{K} \times D}$, where $N_o$ is the number of focal agents and $\mathcal{K}$ is the number of intention queries for each focal agent. To facilitate information interaction among all the intention queries, we reshape them as $E^{(\mathrm{m})}_{\mathrm{I}} \in \mathbb{R}^{N_o\mathcal{K} \times D}$. The process of mutually-guided intention querying can be formulated as follows:

$$\begin{aligned} F'^{(\mathrm{m})}_{\mathrm{I}}[i] = \mathrm{MHSA}\big(&\mathrm{Q}\colon [F^{(\mathrm{m})}_{\mathrm{I}}[i] + E^{(\mathrm{m})}_{\mathrm{I}}[i], \mathrm{PE}(R_{\mathrm{I}[i,i]})], \\ &\mathrm{K}\colon \big\{[F^{(\mathrm{m})}_{\mathrm{I}}[j] + E^{(\mathrm{m})}_{\mathrm{I}}[j], \mathrm{PE}(R_{\mathrm{I}[i,j]})]\big\}_{j\in\Omega(i)}, \\ &\mathrm{V}\colon \big\{F^{(\mathrm{m})}_{\mathrm{I}}[j] + E^{(\mathrm{m})}_{\mathrm{I}}[j] + \mathrm{PE}(R_{\mathrm{I}[i,j]})\big\}_{j\in\Omega(i)}\big), \end{aligned} \tag{2}$$

where $i \in \{1, \ldots, N_o\mathcal{K}\}$, and $F^{(\mathrm{m})}_{\mathrm{I}} \in \mathbb{R}^{(N_o\mathcal{K}) \times D}$ indicates the query content feature from the previous transformer decoder layer and is initialized as zero in the first decoder layer. $R_{\mathrm{I}[i,j]}$ indicate the relative position and direction of $j$-th intention query in the local coordinate system of the $i$-th query token.

Through this information propagation process, the intention queries of multiple agents guide and influence each other during the multimodal motion decoding process. This mutual guidance enhances the capability of the model to generate more informed and realistic predictions for the future trajectories of the agents.

## 3. Experiments

### 3.1. Implementation Details

**Architecture details.** The overall architecture details of MTR++ are the same as MTR [6, 5]. We adopt 6 transformer encoder layers for the context encoding and 6 transformer decoder layers for generating the multimodal future trajectories. The hidden feature dimension is set to 256. For context encoding, the road map is represented as polylines, where each polyline contains up to 20 map points (about $10m$ in WOMD). For the prediction head, a three-layer MLP head is adopted with feature dimension 512. We do not use any traffic light data in our model.

For each category, we adopt 64 intention queries based on 64 intention points that are generated by k-means clustering algorithm on the training set. During testing, we adopt NMS with a distance threshold $2.5m$ to select top 6 predictions from 64 predicted trajectories.

**Training details.** Our model is trained in an end-to-end manner by AdamW optimizer with a learning rate of 0.0001, a weight decay of 0.01, and a batch size of 80 scenes. We train the model for 30 epochs with 8 GPUs, and the learning rate is decayed by a factor of 0.5 every 2 epochs from epoch 20.

**Model ensemble.** To further boost the performance of our MTR++ framework, we adopt the same model ensemble technique as in MTR [6]. Particularly, we trained multiple variants of both MTR and MTR++ models by modifying the

Table 1: Top 10 entries on the test leaderboard of Waymo Open Dataset Motion Prediction Challenge 2023. The Soft mAP is the official ranking metric while the miss rate is the secondary ranking metric. "*" indicates the method is submitted after the deadline of this challenge.

| Method | **Soft mAP**↑ | mAP ↑ | minADE ↓ | minFDE ↓ | Miss Rate ↓ |
|---|---|---|---|---|---|
| MTR++_ens 1st (Ours) | **0.4738** | **0.4634** | **0.5581** | **1.1166** | **0.1122** |
| MTR++ | 0.4414 | 0.4329 | 0.5906 | 1.1939 | 0.1298 |
| *GTR_ens | 0.4518 | 0.4428 | 0.5855 | 1.2056 | 0.1296 |
| IAIR+ 2nd | 0.4480 | 0.4347 | 0.5783 | 1.1679 | 0.1238 |
| GTR-R36 3rd | 0.4384 | 0.4255 | 0.6005 | 1.2225 | 0.1330 |
| GTR | 0.4365 | 0.4230 | 0.5871 | 1.2096 | 0.1309 |
| DM | 0.4362 | 0.4301 | 0.6293 | 1.2723 | 0.1473 |
| MPTr+ | 0.4267 | 0.4130 | 0.5963 | 1.2060 | 0.1318 |
| MPTr | 0.4158 | 0.4018 | 0.6093 | 1.2232 | 0.1336 |
| vtstats | 0.4093 | 0.3976 | 0.6039 | 1.2231 | 0.1364 |
| LeapNet | 0.4089 | 0.3988 | 0.6766 | 1.3203 | 0.1510 |

number of decoder layers, the number of intention queries and the hidden feature dimension. For each focal agent, multiple predictions from these models are first merged together, then we utilize the non-maximum-suppression to select the top 6 predictions based on the predicted confidence of each trajectory.

### 3.2. Main Results

Table 1 presents the leading entries on the final leaderboard of the 2023 Waymo Open Dataset Motion Prediction challenge. The results showcased in Table 1 highlight the outstanding performance of MTR++ even without utilizing model ensemble, already achieving remarkable results on the large-scale Waymo Open Dataset. Upon combining predictions from various MTR++ frameworks, the performance of MTR++ experiences a substantial boost, surpassing all other submissions by a significant margin. These remarkable advancements demonstrate the effectiveness of the proposed MTR++ framework.

## 4. Conclusion

In conclusion, we propose the novel MTR++ framework as a novel solution for the challenges of motion prediction in autonomous driving. Our innovative approaches, including symmetric scene context modeling and mutually-guided intention querying strategies, enhance the representation of scene context information and enable concurrent prediction of multiple agents' trajectories. The results of our experiments demonstrate the outstanding performance of MTR++, leading it to secure the first-place position in the highly competitive Waymo Motion Prediction Challenge 2023. This remarkable achievement contributes to the advancement of autonomous driving systems, offering promising prospects for safer and more efficient transportation in the future.

## References


[1] Junru Gu, Chen Sun, and Hang Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets. In *ICCV*, 2021. 1

[2] Xiaosong Jia, Li Chen, Penghao Wu, Jia Zeng, Junchi Yan, Hongyang Li, and Yu Qiao. Towards capturing the temporal dynamics for trajectory prediction: a coarse-to-fine approach. In *CoRL*, 2023. 1

[3] Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. Multimodal motion prediction with stacked transformers. In *CVPR*, 2021. 1

[4] 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 *ICLR*, 2022. 1

[5] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *NeurIPS*, 2022. 1, 2

[6] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Mtr-a: 1st place solution for 2022 waymo open dataset challenge–motion prediction. *arXiv preprint arXiv:2209.10033*, 2022. 2

[7] Ekaterina Tolstaya, Reza Mahjourian, Carlton Downey, Balakrishnan Vadarajan, Benjamin Sapp, and Dragomir Anguelov. Identifying driver interactions via conditional behavior prediction. In *ICRA*, 2021. 1

[8] 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 *ICRA*, 2022. 1

[9] Maosheng Ye, Tongyi Cao, and Qifeng Chen. Tpcn: Temporal point cloud networks for motion forecasting. In *CVPR*, 2021. 1

[10] 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. In *CoRL*, 2020. 1