Parallel ModeSeq: Technical Report for 2025 Waymo Open Dataset Challenge - Interaction Prediction

Zikang Zhou[1] Haibo Hu[1] Yifan Zhang[2] Yung-Hui Li[3]
Jianping Wang[1] Nan Guan[1] Chun Jason Xue[4]
[1]City University of Hong Kong [2]City University of Hong Kong (Dongguan)
[3]Hon Hai Research Institute [4]Mohamed bin Zayed University of Artificial Intelligence

Abstract

This technical report introduces Parallel ModeSeq, a multimodal behavior prediction framework inspired by sequential mode modeling (ModeSeq). The framework stacks multiple decoding layers with mode rearrangement in between for iterative refinement and applies the Multi-Agent Early-Match-Take-All (MA-EMTA) training strategy to produce multimodal scene output without mode selection or ensemble methods. Our approach efficiently decodes all modes in parallel while translating the mode set into a sequence via causal mode-to-mode self-attention, attaining improved trajectory diversity and calibrated mode confidence without the reliance on recurrent mode generation. We also employ the margin ranking loss to minimize the average number of confidence inversions in ranking, which effectively enhances the performance on mAP. Equipping our framework with the QCNet-style encoding method and the QCNeXt-style joint decoding architecture, our solution achieves phenomenal performance on all metrics, outperforming other methods by a large margin in the 2025 Waymo Interaction Prediction Challenge.

1. Introduction

Multimodal behavior prediction requires capturing the full distribution of agents' future behavior, which is challenging owing to the lack of multimodal ground truth. The problem even becomes exponentially harder when the joint behavior of multiple agents needs to be anticipated. The Waymo Interaction Prediction Challenge [1] has simplified the problem by focusing on two-agent joint prediction, making it feasible to combine marginal prediction results into joint scenes or conduct conditional factorization over agents, both of which necessitate heuristic mode selection (*e.g.*, NMS). However, such approaches are suboptimal, either failing to guarantee scene consistency or violating the end-to-end philosophy of deep learning. Moreover, it is impossible for these approaches to be modified as direct joint prediction models, given that the paradigm of anchor-based dense mode prediction they follow will lead to a prohibitive number of joint anchors. For example, the combination of 64 single-agent anchors from merely two agents will result in 4096 joint anchors, indicating that anchor-based methods are unscalable.

Our solution to the joint prediction problem is fully end-to-end, directly producing a sparse set of multimodal scene outputs without any post-processing tricks. Inspired by ModeSeq [8], we formulate multimodal prediction as sequential mode modeling to obtain diverse trajectories and calibrated confidences under the framework of sparse mode prediction. But unlike ModeSeq, which generates modes in a recurrent manner, our approach leverages causal mode-to-mode self-attention to translate the mode set into a sequence and efficiently decode all modes in one shot, leading to a parallel version of ModeSeq. To further promote the scoring capability, we employ the margin ranking loss to raise the confidence of the positive samples and suppress that of the negative samples. To adapt to the joint multi-agent prediction task, we integrate the parallel version of ModeSeq with the QCNet-style encoder [6] and the QCNeXt-style decoder [7], obtaining a comprehensive prediction framework that achieves unprecedented performance on the Waymo Open Motion Dataset [1].

2. Method

2.1. Scene Encoding

We adopt QCNet [6] as the scene encoder. This encoder exploits a hierarchical map encoding module based on map-to-map self-attention to produce the map embedding of shape $[M, D]$, with D referring to the hidden size. On the other hand, the encoder consists of Transformer modules that factorize the space and time axes, including temporal self-attention, agent-to-map cross-attention, and agent-to-agent self-attention. These three types of attention are grouped and interleaved twice to yield the agent embedding

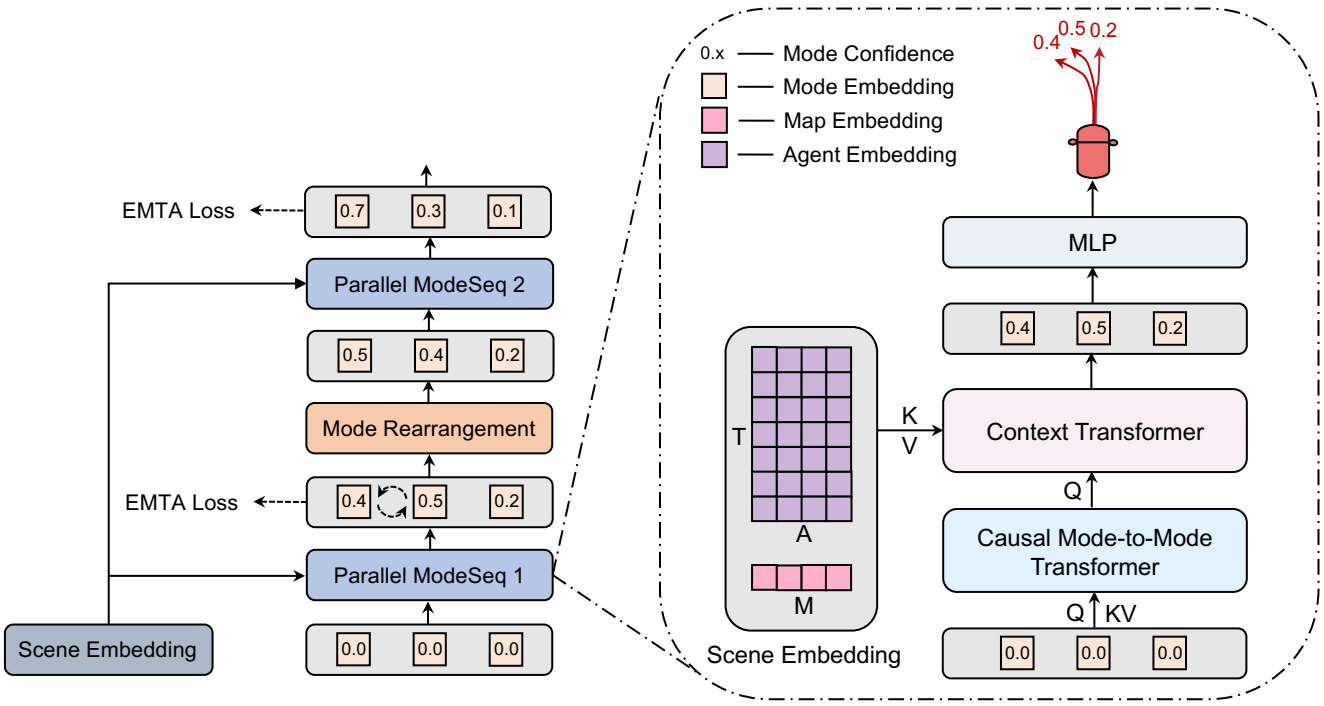

Figure 1. Overview of the Parallel ModeSeq framework for a single agent. **Left:** We stack multiple Parallel ModeSeq layers with mode rearrangement in between to iteratively refine the multimodal output under the Early-Match-Take-All (EMTA) training strategy. **Right:** Each Parallel ModeSeq layer consists of a Causal Mode-to-Mode Transformer module for capturing mode-wise dependencies and a Context Transformer module for retrieving the scene embeddings produced by the encoder.

of shape $[A, T, D]$ for A agents and T historical time steps, which constitutes the final scene embeddings together with the map embedding.

2.2. Marginal Parallel ModeSeq

This section illustrates Parallel ModeSeq for marginal prediction, which can be easily extended as a joint prediction architecture using the modeling tools in QCNeXt [7].

Single-Layer Parallel Mode Sequence. We first introduce the architecture of a single Parallel ModeSeq layer, which is depicted in the right part of Fig. 1. It mainly consists of a causal Mode-to-Mode Transformer and a Context Transformer. The Mode-to-Mode Transformer module captures the relationships between modes, where the query embeddings are derived from the addition of the content embeddings and the order embeddings. The content embeddings are zero-initialized for the first decoding layer. Starting from the second decoding layer, the content embeddings are propagated from the mode embeddings output by the last decoding layer. On the other hand, the order embeddings are learnable parameters used to differentiate the order of modes in the sequence. Unlike typical DETR-like decoders, which treat the query embeddings as set elements, our Mode-to-Mode Transformer module applies causal self-attention, translating the mode set into a sequence by ensuring that the generation of a mode depends on the earlier modes instead of all other modes. Following the Causal Mode-to-Mode Transformer module, we employ a Context Transformer to extract scene context from the encoder embeddings. We decompose the Context Transformer into three separate modules, including mode-time cross-attention, mode-map cross-attention, and mode-agent cross-attention, each of which takes as input only a subset of the encoder embeddings. First, the mode-time cross-attention fuses the query feature with the historical encoding belonging to the agent of interest, enabling the query to adapt to the specific agent. Second, we aggregate the map information surrounding the agent of interest into the query feature leveraging the mode-map cross-attention, which contributes to the map compliance of the forecasting results. Finally, utilizing the mode-agent cross-attention module to fuse the neighboring agents' embeddings at the current time step promotes the model's social awareness. Given the context-aware mode queries, we use MLPs to output the trajectories and the mode confidences.

Multi-Layer Parallel Mode Sequences. Following ModeSeq [8], we use an iterative refinement strategy that stacks multiple Parallel ModeSeq layers and applies training losses to the output of each layer. As shown in the left part of Fig. 1, all layers except for the first one take as input

Table 1. Quantitative results on the 2025 Waymo Open Dataset Interaction Prediction Benchmark.

Method	Ensemble	Soft mAP$_6$ ↑	mAP$_6$ ↑	MR$_6$ ↓	minADE$_6$ ↓	minFDE$_6$ ↓
Parallel ModeSeq	×	**0.2978**	**0.2949**	**0.3782**	**0.7707**	**1.6897**
IMPACT	×	0.2718	0.2659	0.4316	0.9738	2.2734
BeTop	✓	0.2573	0.2511	0.4376	0.9779	2.2805
RetroMotion	✓	0.2562	0.2519	0.4347	0.9256	2.0890

the mode embeddings output from the last round of decoding, refining the features with the scene context. Our model also includes an operation of mode rearrangement in between layers, which corrects the order of the embeddings in the mode sequence to encourage decoding trajectory modes with monotonically decreasing confidence scores. Specifically, before transitioning from the ℓ-th to the $(\ell+1)$-th layer, we sort the mode embeddings according to the descending order of the confidence scores predicted from them. Through iterative refinement with mode rearrangement, the trajectories and the order of modes become more scene-compliant and more monotonous, respectively.

2.3. Extension to Joint Parallel ModeSeq

Thanks to the roto-translation invariance brought by query-centric modeling of QCNet [6], Parallel ModeSeq naturally supports multi-agent prediction in parallel. But this is not enough for the task of joint prediction, which requires scene consistency among the predicted futures of multiple agents. Mirroring the success of QCNeXt [7], we extend the marginal version of Parallel ModeSeq into a joint version in two steps. First, we place a Future Interaction module after the Context Transformer, which performs self-attention among the mode embeddings of all target agents within the same joint scene. Second, we adopt a scene scoring module to produce scene-level confidence scores, where the mode embeddings of all target agents within the same joint scene are aggregated into a scene embedding via max pooling. The scene embedding is then projected into a scene score using an MLP.

2.4. Multi-Agent Early-Match-Take-All training

In this section, we illustrate the training strategy of Joint Parallel ModeSeq, which augments the Early-Match-Take-All (EMTA) scheme [8] into a multi-agent setting, dubbed MA-EMTA. We also employ a margin ranking loss to better calibrate the confidence scores of modes.

Regression. Our regression loss is based on the Laplace negative log-likelihood [5, 6]. Typical WTA loss optimizes only the trajectory with the minimum displacement error to the ground-truth trajectory. In contrast, the MA-EMTA loss optimizes the earliest matched scene in the mode sequence. For example, if both the 2$^{\text{nd}}$ and the 3$^{\text{rd}}$ joint modes match the ground truth, only the 2$^{\text{nd}}$ one will be optimized, regardless of which mode has the minimum displacement error. Here, we decide whether a joint mode is a match based on the velocity-aware distance thresholds defined in the scene-level Miss Rate metric of the Waymo Interaction Prediction Benchmark. If none of the predicted modes match the ground truth, we will fall back to the scene-level WTA loss [4, 7].

Classification. We use the Binary Focal Loss to optimize the confidence scores. As for label assignment, we treat the earliest matches as positive samples, with all the remaining modes treated as negative samples.

Ranking. We desire the confidence scores of the positive samples to be higher than those of the negative samples. To this end, we use the margin ranking loss with a margin of 0.1 to minimize the average number of confidence inversions in each training batch.

3. Experiments

3.1. Implementation Details

We develop models with a hidden size of 128, totaling 11.2M model parameters. The decoder stacks 6 layers for iterative refinement, with each layer predicting 6 modes in parallel. We use the AdamW optimizer [3] to train a single model for 90 epochs on the training set with a batch size of 32, a weight decay rate of 0.1, and a dropout rate of 0.1. The maximum learning rate is set to 5×10^{-4}, which is decayed to 0 every 30 epochs following the cosine annealing schedule with warm restarts [2].

3.2. Quantitative Results

Table 1 shows the quantitative results on the 2025 Waymo Interaction Prediction Leaderboard. Without ensembling, Parallel ModeSeq not only achieves trajectory errors far lower than dense mode prediction methods, but also has much better trajectory diversity and more calibrated scoring capability compared with pseudo joint prediction methods that combine marginal predictions into joint scenes.

4. Conclusion

This technical report presents our solution to the 2025 Waymo Interaction Prediction Challenge. We implement sequential mode modeling in a parallel manner and empower the model with the capability of joint multi-agent prediction, achieving state-of-the-art performance across all prediction metrics.

References

[1] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aurélien Chouard, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *ICCV*, 2021. 1

[2] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 3

[3] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 3

[4] Jiquan Ngiam, Benjamin Caine, Vijay Vasudevan, Zhengdong Zhang, Hao-Tien Lewis Chiang, Jeffrey Ling, Rebecca Roelofs, Alex Bewley, Chenxi Liu, Ashish Venugopal, David Weiss, Ben Sapp, Zhifeng Chen, and Jonathon Shlens. Scene transformer: A unified architecture for predicting multiple agent trajectories. In *ICLR*, 2022. 3

[5] Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *CVPR*, 2022. 3

[6] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *CVPR*, 2023. 1, 3

[7] Zikang Zhou, Zihao Wen, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Qcnext: A next-generation framework for joint multi-agent trajectory prediction. *arXiv preprint arXiv:2306.10508*, 2023. 1, 2, 3

[8] Zikang Zhou, Hengjian Zhou, Haibo Hu, Zihao Wen, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Modeseq: Taming sparse multimodal motion prediction with sequential mode modeling. In *CVPR*, 2025. 1, 2, 3