DeMo: Decoupling Motion Forecasting into Directional Intentions and Dynamic States

Bozhou Zhang **Nan Song** **Li Zhang***
School of Data Science, Fudan University

https://github.com/fudan-zvg/DeMo

*Li Zhang (lizhangfd@fudan.edu.cn) is the corresponding author.

Abstract

Abstract

Accurate motion forecasting for traffic agents is crucial for ensuring the safety and efficiency of autonomous driving systems in dynamically changing environments. Mainstream methods adopt a one-query-one-trajectory paradigm, where each query corresponds to a unique trajectory for predicting multi-modal trajectories. While straightforward and effective, the absence of detailed representation of future trajectories may yield suboptimal outcomes, given that the agent states dynamically evolve over time. To address this problem, we introduce **DeMo**, a framework that decouples multi-modal trajectory queries into two types: mode queries capturing distinct directional intentions and state queries tracking the agent's dynamic states over time. By leveraging this format, we separately optimize the multi-modality and dynamic evolutionary properties of trajectories. Subsequently, the mode and state queries are integrated to obtain a comprehensive and detailed representation of the trajectories. To achieve these operations, we additionally introduce combined Attention and Mamba techniques for global information aggregation and state sequence modeling, leveraging their respective strengths. Extensive experiments on both the Argoverse 2 and nuScenes benchmarks demonstrate that our DeMo achieves state-of-the-art performance in motion forecasting.

The primary distinction between previous methods and ours.

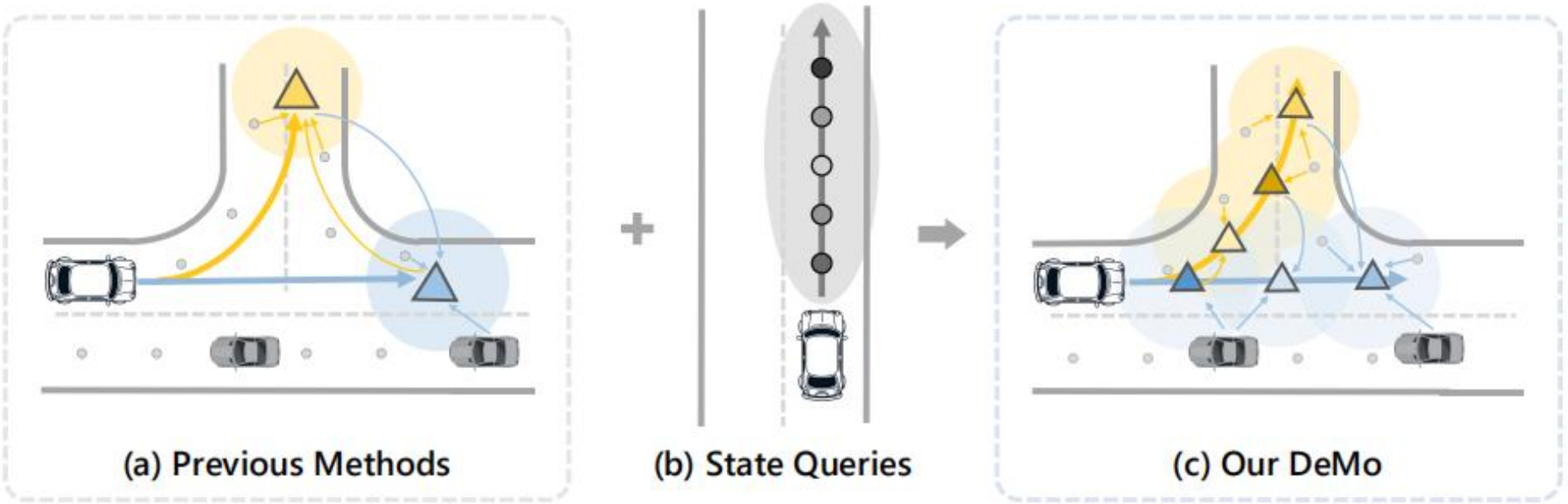

Figure 1: The primary distinction between previous methods and ours lies in the representation of future trajectories. Previous methods, as depicted in (a), use only one mode query for each trajectory. Our approach, illustrated in (c), utilizes decoupled mode queries, as shown in (a), and state queries, as shown in (b), to represent the multi-modal trajectories.

Overview of our DeMo framework.

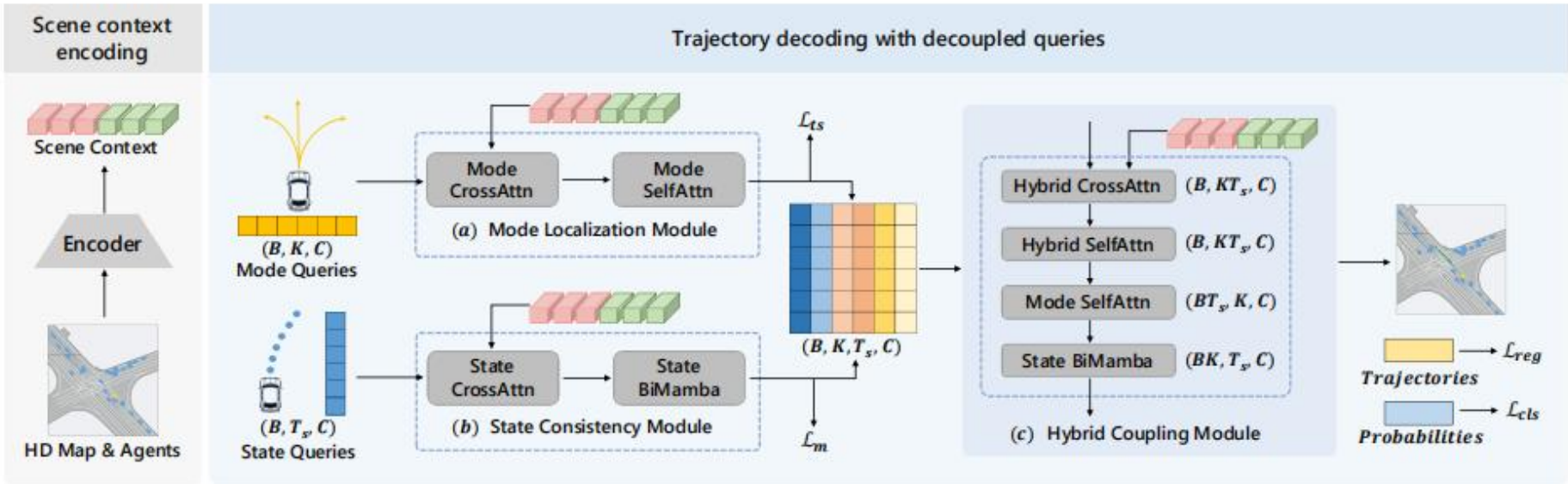

Figure 2: Overview of our DeMo framework: The HD maps and agents are first processed by the encoder to obtain the scene context. The decoding pipeline includes: (a) the Mode Localization Module, which processes mode queries by interacting with the scene context from the encoder and among themselves; (b) the State Consistency Module, which processes state queries; and (c) the Hybrid Coupling Module, which combines these queries to generate the final output. The feature dimension is illustrated using a single-agent setting, where B represents the batch size.

Experiment on Argoverse 2.

Table 1: Performance comparison on *Argoverse 2 single-agent test set* in the official leaderboard. For each metric, the best result is in **bold** and the second best result is underlined. The upper part features a single model, while the lower part employs model ensembling as a trick.

Method	$minFDE_1$	$minADE_1$	$minFDE_6$	$minADE_6$	MR_6	$b\text{-}minFDE_6$
FRM [47]	5.93	2.37	1.81	0.89	0.29	2.47
HDGT [31]	5.37	2.08	1.60	0.84	0.21	2.24
SIMPL [72]	5.50	2.03	1.43	0.72	0.19	2.05
THOMAS [22]	4.71	1.95	1.51	0.88	0.20	2.16
GoRela [11]	4.62	1.82	1.48	0.76	0.22	2.01
MTR[54]	4.39	1.74	1.44	0.73	0.15	1.98
HPTR [73]	4.61	1.84	1.43	0.73	0.19	2.03
GANet [64]	4.48	1.77	1.34	0.72	0.17	1.96
ProphNet [65]	4.74	1.80	1.33	0.68	0.18	1.88
QCNet [77]	4.30	1.69	1.29	0.65	0.16	1.91
SmartRefine [76]	4.17	1.65	1.23	0.63	0.15	1.86
DeMo (Ours)	**3.74**	**1.49**	**1.17**	**0.61**	**0.13**	**1.84**
QML [57]	4.98	1.84	1.39	0.69	0.19	1.95
TENET [66]	4.69	1.84	1.38	0.70	0.19	1.90
MacFormer [15]	4.69	1.84	1.38	0.70	0.19	1.90
BANet [70]	4.61	1.79	1.36	0.71	0.19	1.92
Gnet [19]	4.40	1.72	1.34	0.69	0.18	1.90
Forecast-MAE [8]	4.15	1.66	1.34	0.69	0.17	1.91
QCNet [77]	3.96	1.56	1.19	0.62	0.14	1.78
DeMo (Ours)	**3.70**	**1.49**	**1.11**	**0.60**	**0.12**	**1.73**

Experiment on nuScenes.

Table 2: Performance comparison on *nuScenes test set* in the official leaderboard. The "-" symbol means the corresponding metric is unknown.

Method	$minFDE_1$	$minADE_5$	$minADE_{10}$	MR_5	MR_{10}
Trajectron++ [51]	9.52	1.88	1.51	0.70	0.57
LaPred [33]	8.37	1.47	1.12	0.53	0.46
P2T [13]	10.50	1.45	1.16	0.64	0.46
GOHOME [21]	6.99	1.42	1.15	0.57	0.47
CASPNet [52]	-	1.41	1.19	0.60	0.43
Autobot [23]	8.19	1.37	1.03	0.62	0.44
THOMAS [22]	<u>6.71</u>	1.33	1.04	0.55	<u>0.42</u>
PGP [14]	7.17	1.27	<u>0.94</u>	0.52	**0.34**
LAformer [39]	6.95	**1.19**	1.19	<u>0.48</u>	0.48
DeMo (Ours)	**6.60**	<u>1.22</u>	**0.89**	**0.43**	**0.34**

Ablation study.

Table 4: Ablation study on the core components of DeMo on the *Argoverse 2 single-agent validation set*. "Decpl. Query" indicates decoupled query paradigm. "Agg. Module" indicates three aggregation modules. "Aux. Loss" indicates two auxiliary losses.

ID	State Query	Decpl. Query	Agg. Module	Aux. Loss	$minFDE_1$	$minADE_1$	$minFDE_6$	$minADE_6$	MR_6	$b\text{-}minFDE_6$
1					4.489	1.792	1.414	0.750	0.184	2.067
2	✓				4.494	1.800	1.505	0.777	0.208	2.138
3	✓	✓		✓	4.385	1.746	1.405	0.761	0.180	2.051
4	✓	✓	✓		4.247	1.695	1.319	0.687	0.166	1.961
5	✓	✓	✓	✓	**3.917**	**1.609**	**1.268**	**0.674**	**0.152**	**1.918**

Qualitative results.

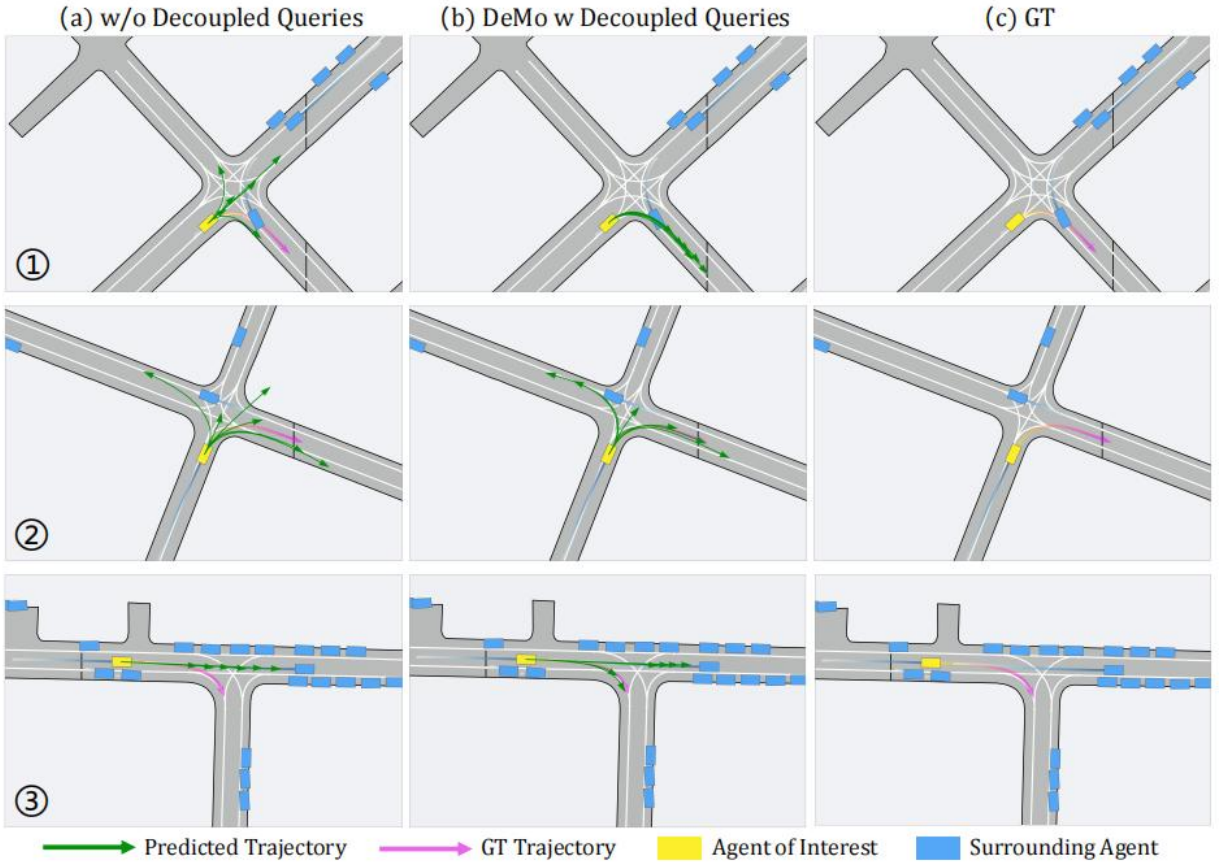

Figure 3: Qualitative results on the Argoverse 2 single-agent validation set. Panel (a) illustrates the results of the baseline model without decoupled queries; Panel (b) illustrates the results of our DeMo, which employs decoupled queries; and Panel (c) represents the ground truth.

Thank you for listening!