MGTR: Multi-Granular Transformer for Motion Prediction with LiDAR

Yiqian Gan*, Hao Xiao*, Yizhe Zhao*, Ethan Zhang, Zhe Huang, Xin Ye, Lingting Ge

Contact: Yiqian Gan gan0913@gmail.com Hao Xiao alexinsjtu@gmail.com

This work was done at TuSimple Inc.

ICRA2024

YOKOHAMA | JAPAN

Abstract

Motion prediction has been an essential component of autonomous driving systems since it handles highly uncertain and complex scenarios involving moving agents of different types. In this paper, we propose a Multi-Granular TRansformer (MGTR) framework, an encoder-decoder network that exploits context features in different granularities for different kinds of traffic agents. To further enhance MGTR's capabilities, we leverage LiDAR point cloud data by incorporating LiDAR semantic features from an off-the-shelf LiDAR feature extractor. We evaluate MGTR on Waymo Open Dataset motion prediction benchmark and show that the proposed method achieved state-of-the-art performance, ranking 1st on its leaderboard*.

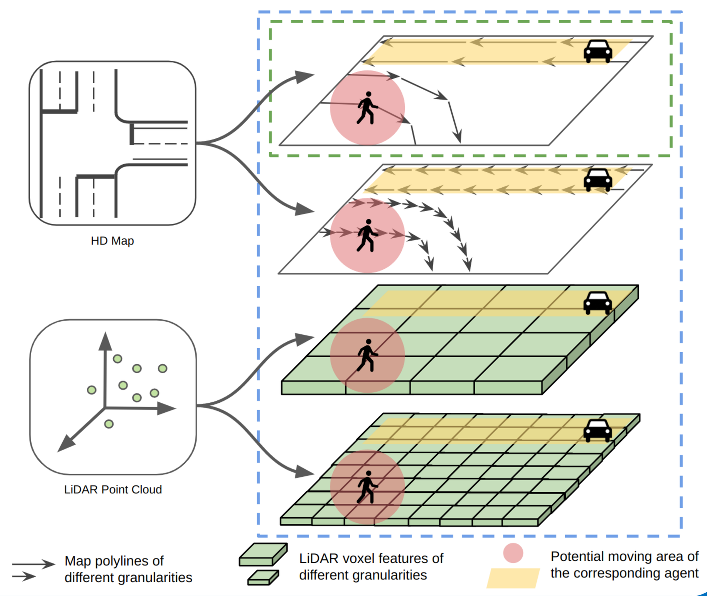
* https://waymo.com/open/challenges/2023/motion-prediction/

Motivation

Most previous methods encode road graph only in a single granularity for all agents in the scene (green dashed box). In our method, various agents can benefit from multi-granular context information encoded from multimodal sources (blue dashed box).

Motivation 1: The needs of granularity from different types of agents (e.g. vehicles and pedestrians) are different.

Motivation 2: LiDAR, serving as a dense online perception representation, is able to provide aforementioned context information to improve prediction performance.



Method

Multimodal input, including agents, HD map and LiDAR.

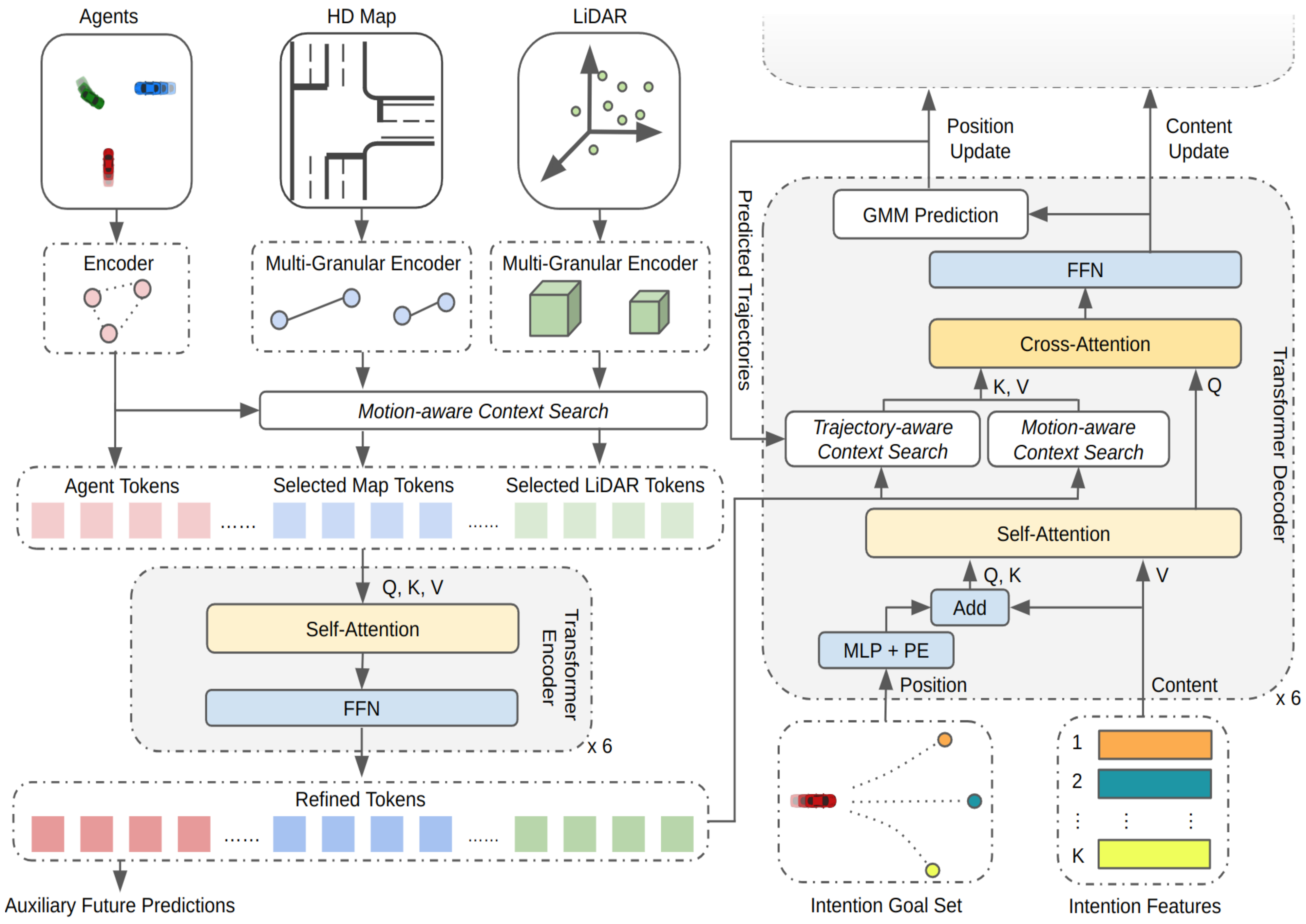
Multi-Granular context encoder encodes HD map and LiDAR in multiple granularities.

$$F_A = \phi\big(\mathrm{MLP}(\Gamma(P_A))\big), \quad F_M^{(i)} = \phi\big(\mathrm{MLP}(\Gamma(P_M^{(i)}))\big),$$
$$F_L^{(i)} = \mathrm{MLP}\big(\mathcal{P}^{(i)}(\Gamma(\mathcal{V}))\big),$$

Motion-aware context search is introduced to select more meaningful context for agents with different motion patterns.

Transformer encoder:

- **Token aggregation and encoding** with local self attention.
  $$Q = F_e^{j-1} + PE(F_e^{j-1}), \quad V = \kappa(F_e^{j-1}),$$
  $$K = \kappa(F_e^{j-1}) + PE(\kappa(F_e^{j-1})),$$
  $$F_e^j = \mathrm{MHSA}(Q, K, V),$$
- **Future state enhancement**: a future trajectory is predicted for each agent and it can be formulated as $\mathcal{T}_{scene} = \mathrm{MLP}(F_e^A)$,



Losses

- Auxiliary task loss on future predicted trajectories,
- Classification loss on predicted intention probability,
- GMM loss in form of negative log-likelihood loss of the predicted trajectories.

Updated position and content feature will be used by the next decoder layer.

Transformer decoder:

- **Self-attention** to propagate information among K intention queries.
  $$Q = K = F_d^{j-1} + PE(F_d^{j-1}), \quad V = F_d^{j-1},$$
  $$F_I^j = \mathrm{MHSA}(Q, K, V),$$
- **Cross-attention** to aggregate features from refined tokens.
  $$Q = F_I^j + PE(F_I^j),$$
  $$K = V = \gamma(F_e) + PE(\gamma(F_e)),$$
  $$F_d^j = \mathrm{MHCA}(Q, K, V),$$
  $$\gamma(F_e) = \eta(F_e) \cup \theta(F_e),$$
- **Gaussian Mixture Model** to represent K trajectories corresponding to K intention queries.

K representative intention goals are used. Each intention goal represents an implicit motion mode.

Experiments

TABLE I

COMPARISON ON WOMD-LIDAR VAL SET. RESULTS IN TOP THREE ROWS ARE COMPUTED AS AN AVERAGE OF $t = 3, 5,$ AND 8 SECONDS, WHILE THE ONES IN BOTTOM THREE ROWS ARE REPORTED FOR $t = 8$ SECONDS. MTR++ [23] DOES NOT REPORT CATEGORICAL RESULTS. FOLLOWING [10], ALL METRICS ARE REPORTED WITH TWO DECIMAL PLACES. * INDICATES METHODS UTILIZING LIDAR.

	mAP ↑			
Method	Vehicle	Pedestrian	Cyclist	Average
MTR [6]	0.45	0.44	0.36	0.42
MTR++ [23]	-	-	-	0.44
MGTR* (Ours)	**0.46**	**0.47**	**0.40**	**0.45**
Wayformer [22]	0.35	0.35	0.29	0.33
Wayformer+LiDAR [10]*	0.37	0.37	0.28	0.34
MGTR* (Ours)	**0.38**	**0.44**	**0.32**	**0.38**

Compare with MTR (1st of 2022 Waymo Open Dataset Challenge) and MTR++* (1st of 2023 Waymo Open Dataset Challenge) on average mAP of t = 3s, 5s and 8s, we show a non-trivial improvement across all category.

Compare with Wayformer and Wayformer+LiDAR (The only multimodal model with LiDAR input) on mAP of t = 8s*, we show a whopping 7% improvement in terms of mAP on pedestrian category.

TABLE II

COMPARISON ON WOMD-LIDAR TEST SET. ALL METRICS ARE AVERAGED OVER 3S, 5S, AND 8S. ALL MODELS DO NOT USE MODEL ENSEMBLE.

	Vehicle				Pedestrian				Cyclist				Avg
Method	minADE↓	minFDE↓	MR↓	mAP↑	minADE↓	minFDE↓	MR↓	mAP↑	minADE↓	minFDE↓	MR↓	mAP↑	mAP↑
ReCoAt [39]	0.9865	2.1771	0.2695	0.2667	0.4261	0.8982	0.1451	0.3208	0.8985	1.9252	0.3164	0.2258	0.2711
DenseTNT [19]	1.3462	1.9120	0.1518	0.3698	0.5013	0.9130	0.1014	0.3342	1.2687	1.8292	0.2186	0.2802	0.3281
SceneTransformer [21]	**0.7094**	**1.4115**	0.1480	0.3270	0.3812	0.7532	0.0971	0.2715	0.7446	1.4701	0.2239	0.2380	0.2788
GTR-R36 [40]	0.7450	1.5049	0.1477	0.4521	0.3470	0.7221	0.0741	0.4243	0.7095	1.4406	0.1772	0.4003	0.4255
DM [41]	0.7701	1.5400	0.1529	0.4725	0.3741	0.7882	0.0848	0.4172	0.7436	1.4885	0.2043	0.4005	0.4301
MTR [6]	0.7642	1.5257	0.1514	0.4494	0.3486	0.7270	0.0753	0.4331	0.7022	**1.4093**	0.1786	0.3561	0.4129
MTR++ [23]	0.7178	1.4321	**0.1366**	**0.4871**	0.3504	0.7305	0.0745	0.4324	0.7036	1.4190	0.1784	0.3792	0.4329
MGTR (Ours)	0.7393	1.5119	0.1497	0.4626	**0.3441**	**0.7191**	**0.0722**	**0.4865**	**0.6919**	1.4096	**0.1675**	**0.4023**	**0.4505**

Compared to the latest SOTA motion prediction model, MTR++, we achieve a whopping 5.41% increase in terms of mAP on the pedestrian category. This strongly signals that for non-vehicular objects, features that attend to details are key to more accurate trajectory predictions.

TABLE III

ABLATION STUDY ON OUR PROPOSED MGTR.

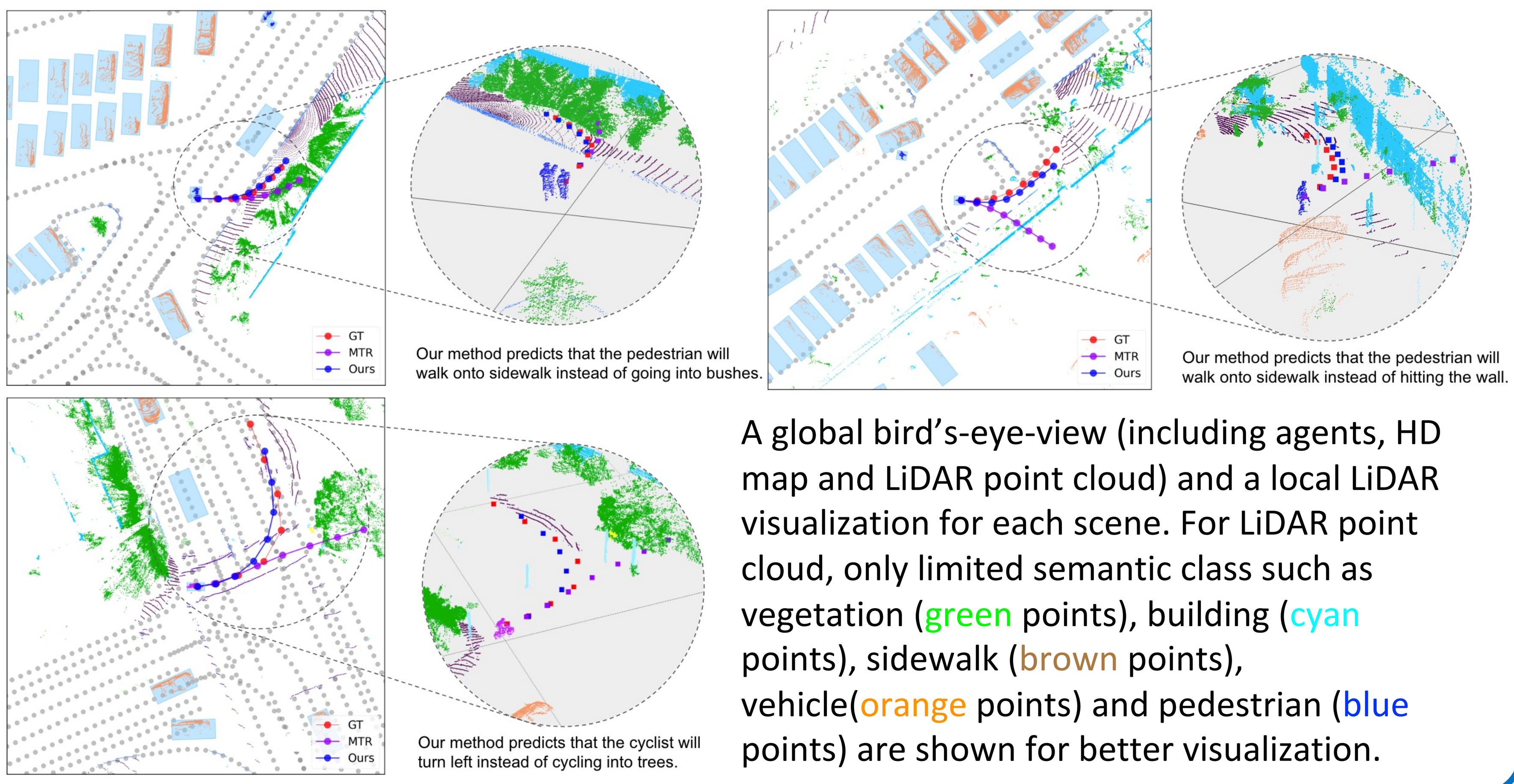
	mAP ↑			
Description	Vehicle	Pedestrian	Cyclist	Average
Baseline	0.3860	0.3682	0.2881	0.3474
+ multi-granular map	0.3895	0.3730	0.2900	0.3508
+ multi-granular LiDAR	0.3896	0.3820	0.2997	0.3571
+ motion-aware context search	**0.3919**	**0.3935**	**0.3025**	**0.3626**

Method Name	Soft mAP ↓	mAP	minADE	minFDE	Miss rate	Overlap Rate
MGTR_ens	0.4764	0.4658	0.5825	1.2009	0.1258	0.1270
MTR++_Ens	0.4738	0.4634	0.5581	1.1166	0.1122	0.1276
MGTR	0.4599	0.4505	0.5918	1.2135	0.1298	0.1275
GTR_ens	0.4518	0.4428	0.5855	1.2056	0.1296	0.1277
EDA_single	0.4510	0.4401	0.5718	1.1702	0.1169	0.1266
IAIR+	0.4480	0.4347	0.5783	1.1679	0.1238	0.1263
AMP	0.4451	0.4334	0.6021	1.1918	0.1311	0.1388
ControlMTR	0.4436	0.4285	0.5904	1.1964	0.1295	0.1260
MTR++	0.4414	0.4329	0.5906	1.1939	0.1298	0.1281

Our ensembled model (MGTR_ens)

Our Single model (MGTR)

Visualization



A global bird's-eye-view (including agents, HD map and LiDAR point cloud) and a local LiDAR visualization for each scene. For LiDAR point cloud, only limited semantic class such as vegetation (green points), building (cyan points), sidewalk (brown points), vehicle(orange points) and pedestrian (blue points) are shown for better visualization.

References

[6] S. Shi, L. Jiang, D. Dai, and B. Schiele, "Motion transformer with global intention localization and local movement refinement," Advances in Neural Information Processing Systems, vol. 35, pp. 6531–6543, 2022.

[10] K. Chen, R. Ge, H. Qiu, R. Ai-Rfou, C. R. Qi, X. Zhou, Z. Yang, S. Ettinger, P. Sun, Z. Leng, M. Mustafa, I. Bogun, W. Wang, M. Tan, and D. Anguelov, "Womd-lidar: Raw sensor dataset benchmark for motion forecasting," 2023.

[21] J. Ngiam, B. Caine, V. Vasudevan, Z. Zhang, H.-T. L. Chiang, J. Ling, R. Roelofs, A. Bewley, C. Liu, A. Venugopal et al., "Scene transformer: A unified architecture for predicting multiple agent trajectories," arXiv preprint arXiv:2106.08417, 2021.

[22] N. Nayakanti, R. Al-Rfou, A. Zhou, K. Goel, K. S. Refaat, and B. Sapp, "Wayformer: Motion forecasting via simple & efficient attention networks," in 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023, pp. 2980–2987.

[23] S. Shi, L. Jiang, D. Dai, and B. Schiele, "Mtr++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying," arXiv preprint arXiv:2306.17770, 2023.