ICRA2024
YOKOHAMA | JAPAN

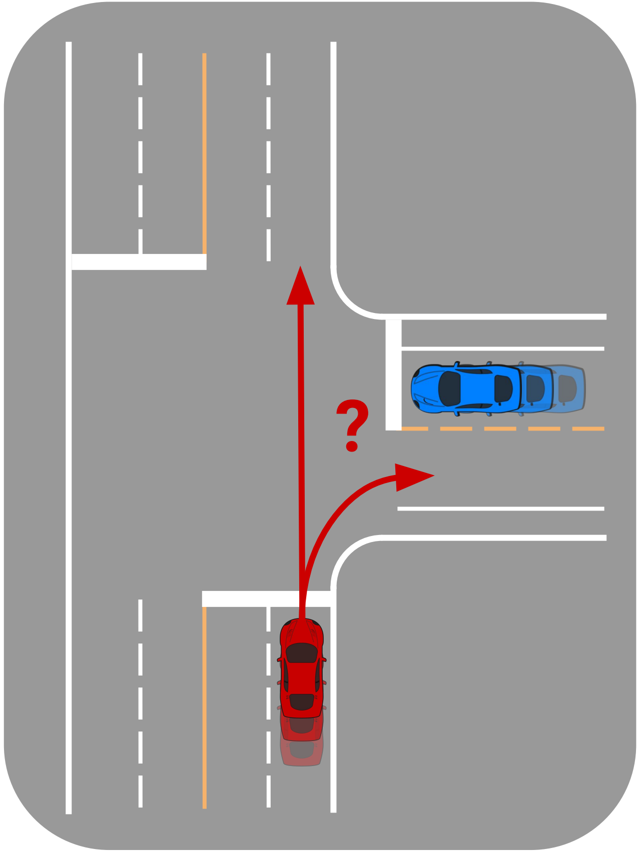
MGTR: Multi-Granular Transformer for Motion Prediction with LiDAR

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

tu simple

MGTR

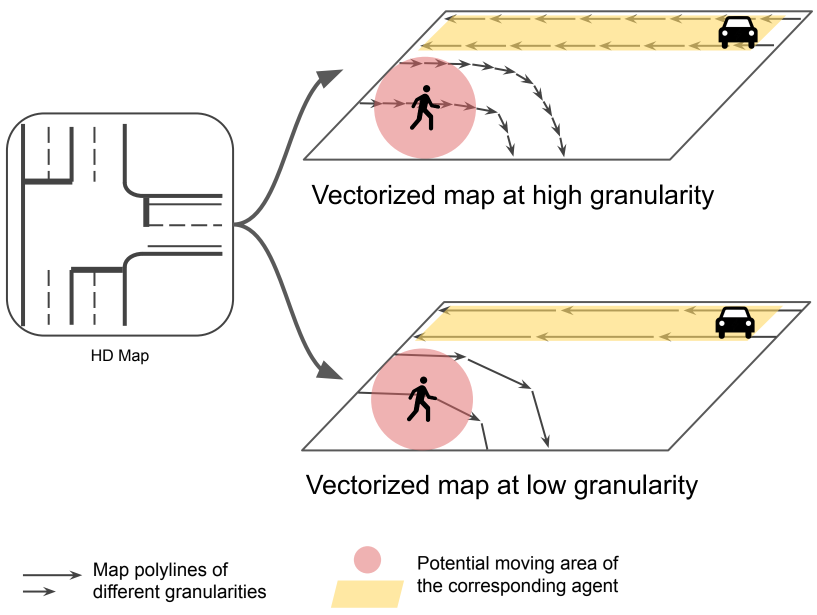
- **Multi-granular context encoding** is integrated in a Transformer-based motion prediction framework for the first time.
- **LiDAR** is introduced as an additional rich 3D context information to overcome limitations of pre-built HD maps.
- **SOTA** performance has been achieved on Waymo motion prediction dataset (**rank 1st** as of the paper submission date).



Motion Prediction Visualization

Motivation-1

Single-granular context encoding is not enough for all agents.



Map context as an example, during map vectorization, if with same number of map tokens:

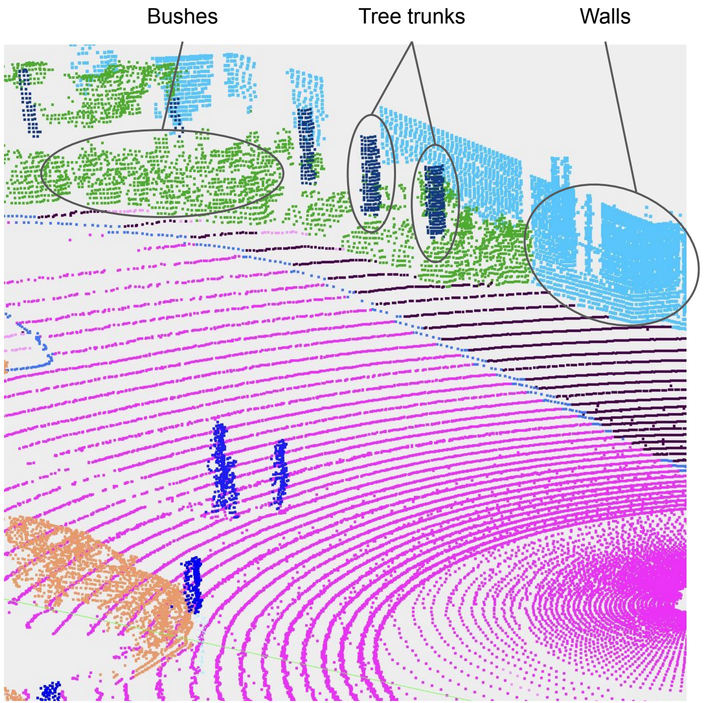
1. High-granularity encoding generates map vectors that providing more fine-grained road structure information (e.g., curvature).

2. Low-granularity map provides a larger perception range but lower resolution

Distinct motion patterns of different types of agents result in different needs of granularity (e.g. vehicles and pedestrians). Therefore, it's beneficial to make multi-granular context accessible to every agent.

Motivation-2

Using only pre-built maps as context input has limitations that can be addressed by LiDAR.



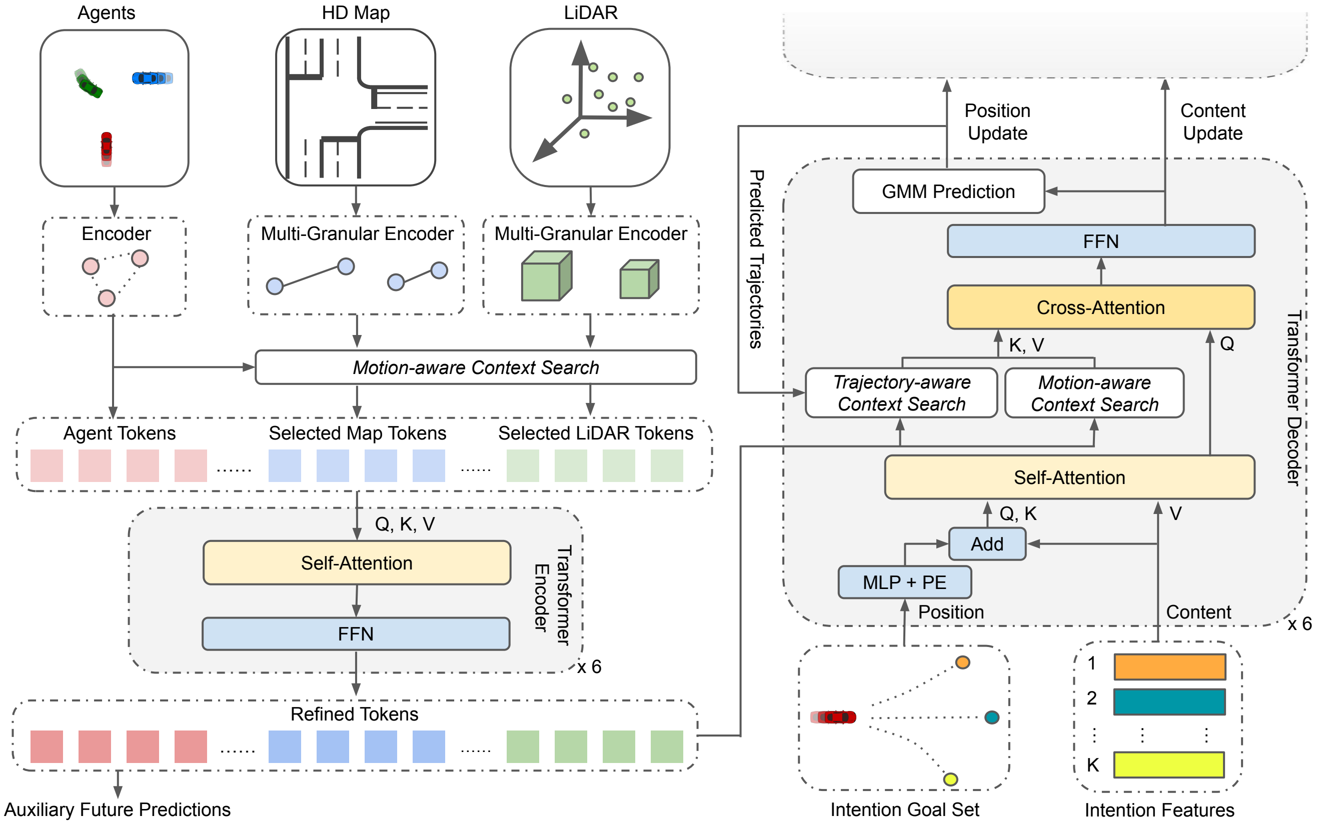
LiDAR Semantic Segmentation Visualization

At least two types of important context information are missing from traditional pre-built maps:

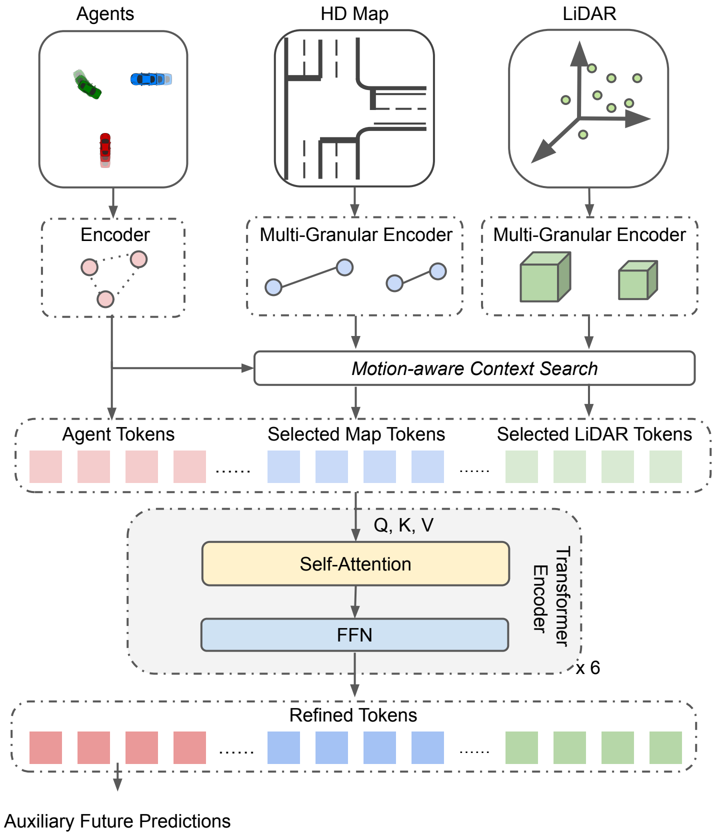
1. Uncountable amorphous regions that are hard to represent as instances in maps, such as road verges, bushes and walls.
2. Temporary road structures that are not included in maps, such as temporary traffic cones and construction zones.

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

Framework: An overview

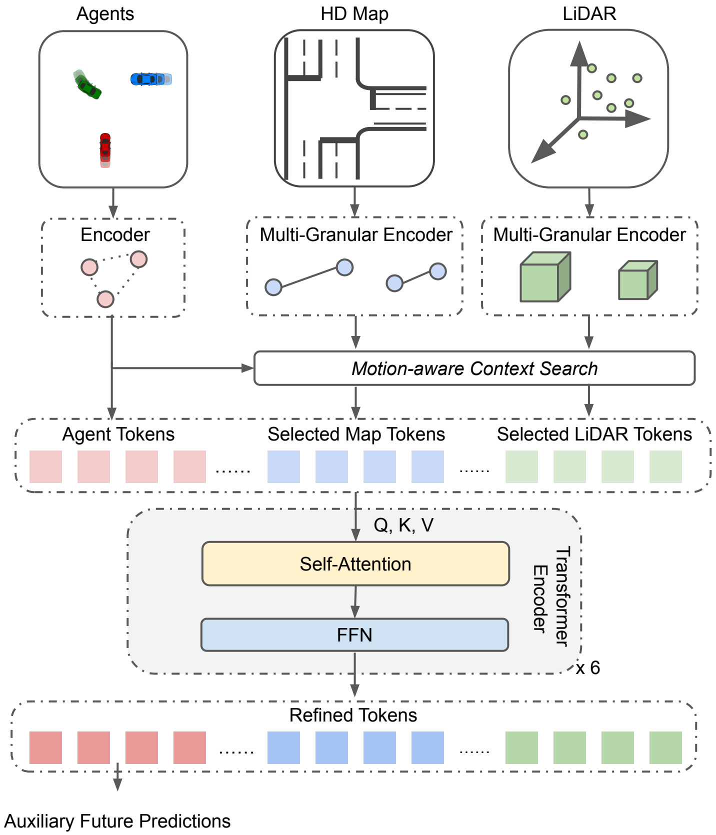


Framework: Transformer encoder



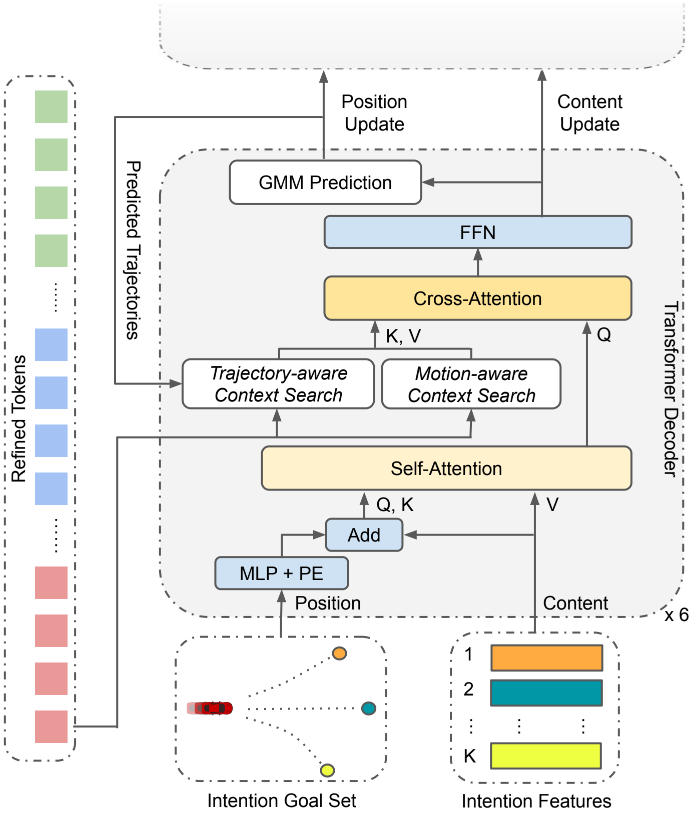
- **Multimodal input**, including agents, HD map and LiDAR.
- **Multi-Granular context encoder** encodes HD map and LiDAR in multiple granularities.
- **Motion-aware context search** is introduced to select more meaningful context for agents with different motion patterns.
- **Local self attention** is adopted to aggregate information from multimodal multi-granular tokens.

Framework: Transformer encoder



- **Multimodal input**, including agents, HD map and LiDAR.
- **Multi-Granular context encoder** encodes HD map and LiDAR in multiple granularities.
- **Motion-aware context search** is introduced to select more meaningful context for agents with different motion patterns.
- **Local self attention** is adopted to aggregate information from multimodal multi-granular tokens.
- **Auxiliary future motion prediction** task is added to further improve encoding performance

Framework: Transformer decoder



- Update positions and content feature to the next decoder layer.

- In the Transformer decoder, we apply:
  - Self-attention to propagate information among K intention queries.
  - Trajectory-aware and motion-aware context search
  - Cross-attention to aggregate features from refined tokens.
  - 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.

Quantitative Result - Validation Set

TABLE I

Method	mAP ↑ 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 categories.

*MTR++ does not report its categorical result on validation.

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 and 4% on cyclist category.

*Wayformer series only report their mAP of 8s instead of the average of 3s, 5s and 8s. To be fair, we compare our method with Wayformer series on mAP of 8s as well.

Quantitative Result - Test Set

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**

This strongly signals that for non-vehicular objects, features that attend to details are key to more accurate and reliable trajectory predictions.

Waymo Open Dataset leaderboard - Motion Prediction*

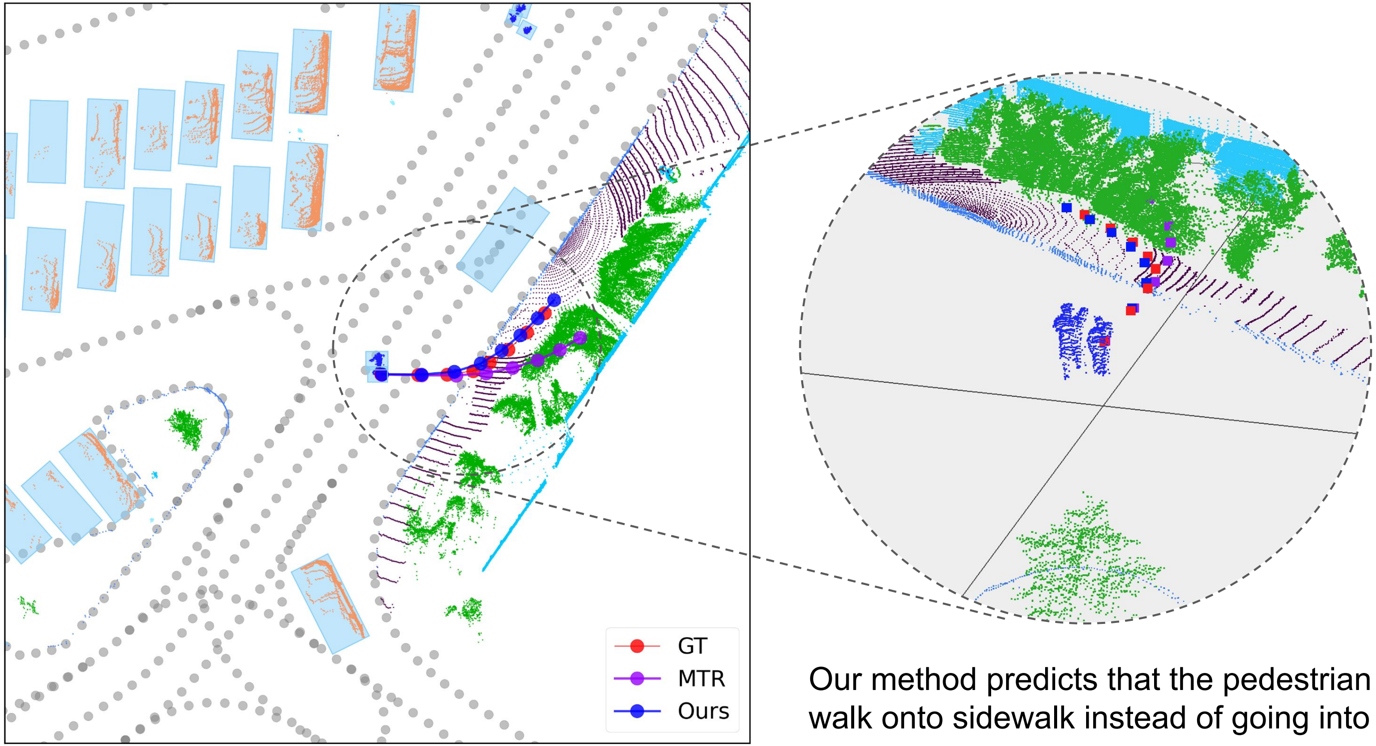
Method Name	Lidar data for training	Object Type	Evaluation Time	Soft mAP	mAP	minADE	minFDE	Miss Rate	Overlap Rate	Date (Pacific Daylight Time)
		All	Avg	Show rest						
MGTR_ens		All	Avg	0.4764	0.4658	0.5825	1.2009	0.1258	0.1270	2023-09-15 19:06
MTR++_Ens		All	Avg	0.4738	0.4634	0.5581	1.1166	0.1122	0.1276	2023-05-23 15:37
MGTR		All	Avg	0.4599	0.4505	0.5918	1.2135	0.1298	0.1275	2023-09-14 21:18
GTR_ens		All	Avg	0.4518	0.4428	0.5855	1.2056	0.1296	0.1277	2023-05-25 02:58
EDA_single		All	Avg	0.4510	0.4401	0.5718	1.1702	0.1169	0.1266	2023-08-07 07:23
IAIR+		All	Avg	0.4480	0.4347	0.5783	1.1679	0.1238	0.1263	2023-05-23 23:56
MTR++		All	Avg	0.4414	0.4329	0.5906	1.1939	0.1298	0.1281	2023-05-23 12:31
GTR-R36		All	Avg	0.4384	0.4255	0.6005	1.2225	0.1330	0.1279	2023-05-23 20:39
GTR		All	Avg	0.4365	0.4230	0.5871	1.2096	0.1309	0.1272	2023-05-16 17:50
DM		All	Avg	0.4362	0.4301	0.6293	1.2723	0.1473	0.1364	2023-05-23 23:39

Our ensembled model

Our single model

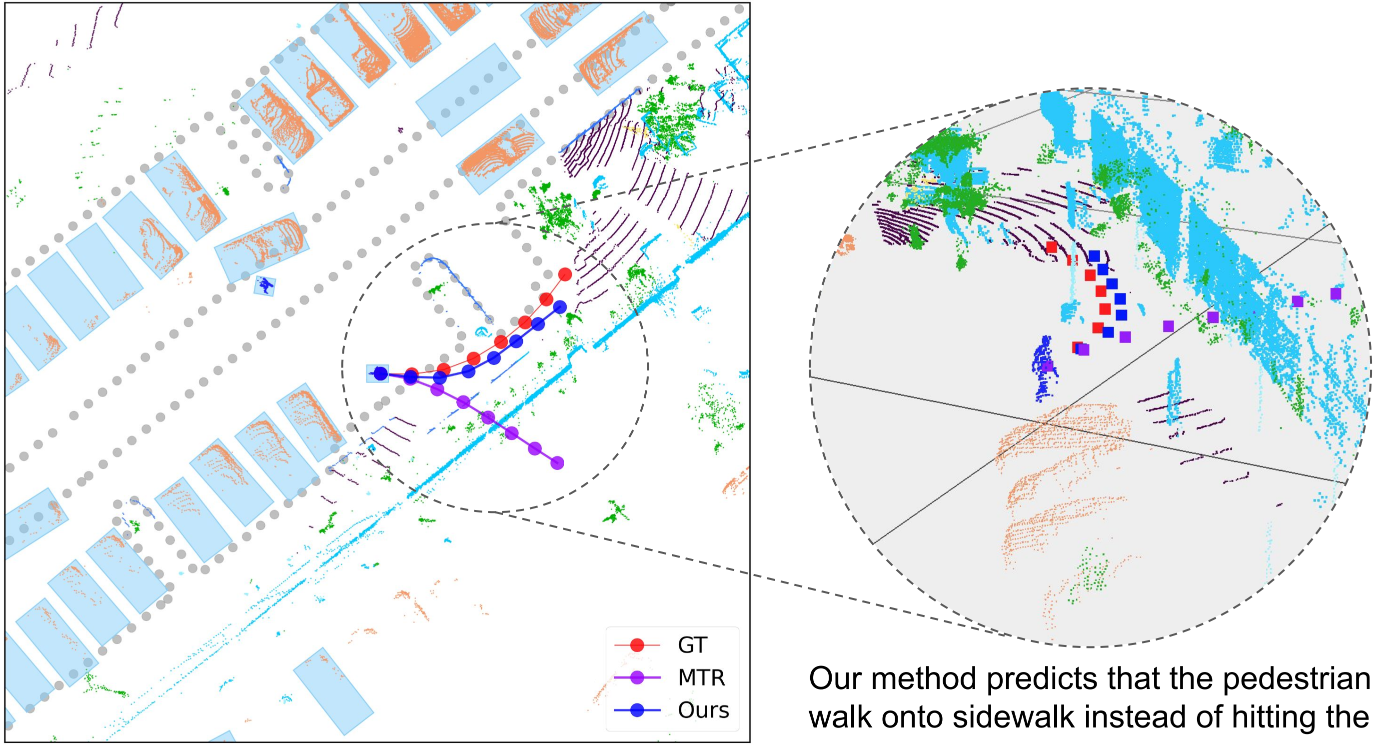
*Top 10 entries on the leaderboard of motion prediction track of Waymo Open Dataset, which includes both single model results and ensemble model results.

Qualitative Result 1



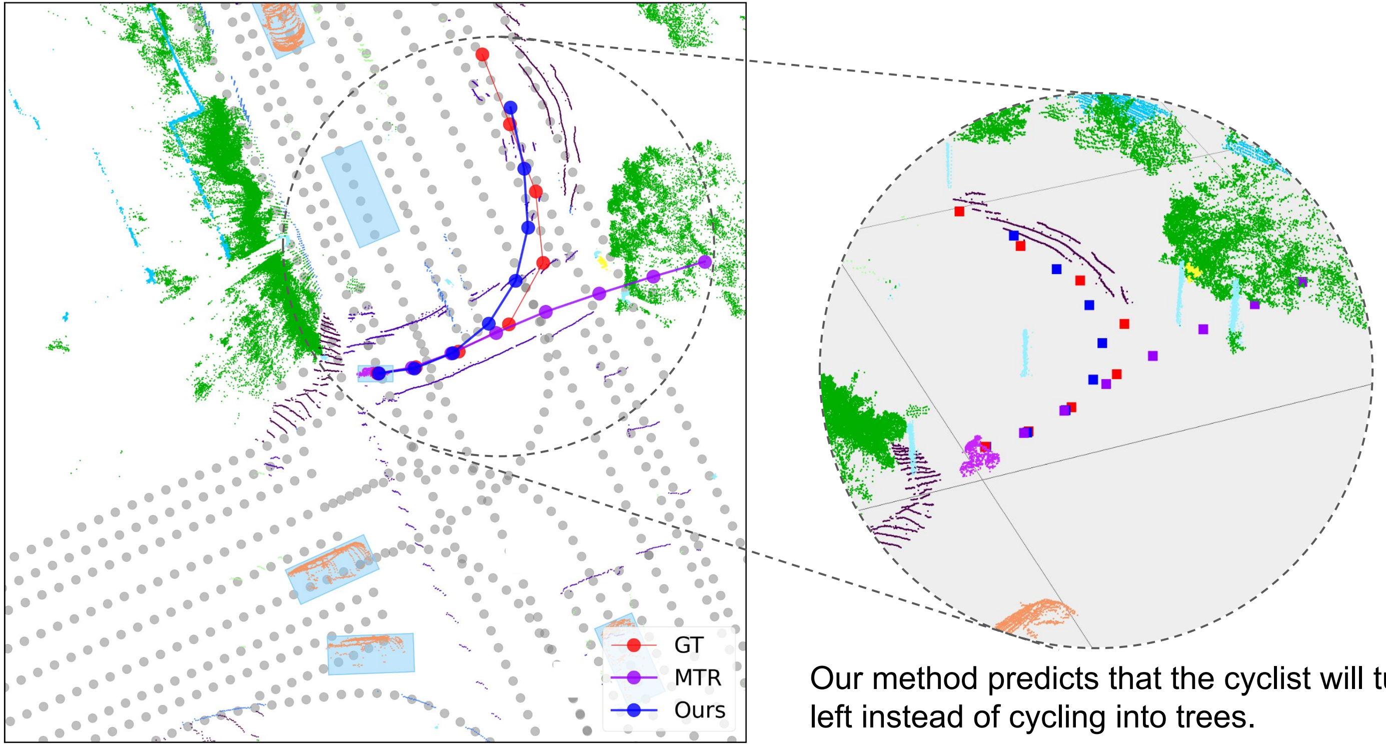
Our method predicts that the pedestrian will walk onto sidewalk instead of going into bushes.

Qualitative Result 2



Our method predicts that the pedestrian will walk onto sidewalk instead of hitting the wall.

Qualitative Result 3



Our method predicts that the cyclist will turn left instead of cycling into trees.

References

- 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.
- 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.
- 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.
- 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.
- 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.

ICRA2024
YOKOHAMA | JAPAN

Thank You For Listening!

Further reachout:

Yiqian Gan: gan0913@gmail.com

Hao Xiao: alexinsjtu@gmail.com

Future Directions

- Model capacity and efficiency
- Adapt other types of sensor inputs such as Radar, Camera images
- Integration with upper stream tasks such as segmentation, detection, tracking, fusion etc.
- Integration with the planning task and enrich context reasoning.

Training & Losses

MGTR employs a weighted combination of losses including

- 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