Transformer with Group-wise Modal Assignments for Motion Prediction

Haochen Liu, Xiaoyu Mo, Zhiyu Huang, Chen Lv
Nanyang Technological University, Singapore
haochen002@e.ntu.edu.sg, xiaoyu.mo@ntu.edu.sg, zhiyu001@e.ntu.edu.sg, lyuchen@ntu.edu.sg

Abstract

*Accurately predicting the trajectory of target traffic actors with uncertainty awareness is paramount for improving the safety and interaction capabilities of autonomous vehicles, especially in complex scenarios. Addressing aleatoric uncertainty through a multi-modal paradigm presents a remaining challenge, which involves reconciling the granularity of modalities with the limited number of positive training samples allocated for each modality. This work proposes **GTR-R36**: a Motion Transformer (MTR) backboned motion predictor with group-wise modal allocation strategy in trajectory decoding. To tackle this challenge, our approach involves several key steps. Firstly, we devised group-wise querying modalities with unique initialization for each group, thereby enhancing the diversity and quantity of positive samples for each respective modality. Additionally, we developed a miss-rate optimization scheme that identifies and delineates the hit region from the missed ones across group queries within each modality. Through our proposed method, we achieved competitive prediction accuracy and demonstrated impressive performance across various evaluation metrics in the WOMD benchmark.*

1. Introduction

Making accurate motion forecasting of targeted traffic participants (actors) poses one of the most formidable challenges in the realm of autonomous driving [1]. Notably, it presents a highly challenging task due to the hereditary uncertain behaviors, or multi-modalities for each prediction target under intricate driving environments [7]. To properly address this challenge, existing approaches are primarily twofold involving goal-based estimations and set-based regressions with a mixture of modalities. The former approaches tackle the uncertainty conditioning on an assumption of goal prior for subsequent trajectory predictor [3, 4], while the latter conduct direct update of the trajectory predictor with modality scoring and binning by certain criteria [5, 6]. Still, goal-based methods encounters significant computational burdens as their performance relies on considering a vast number of goal candidates, and may require massive iterations to sample for the targeting goals.

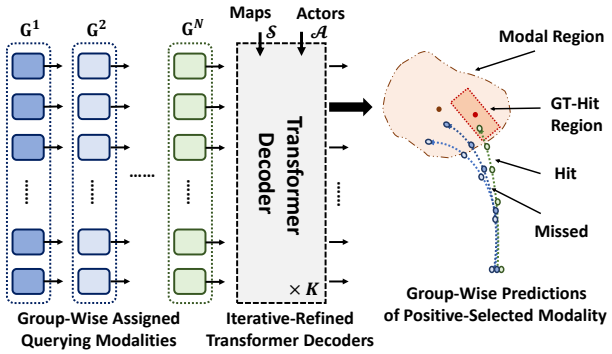

Figure 1. **Method overview:** N Groups of assigned querying modalities $\mathbf{G}^N$ are initialized and simultaneously decoded using shared Transformer structures. Miss-rate optimization is performed for predict positive-selected modality across these groups, effectively delineating the ground-truth hit region from the rest of heuristic modal region for each modality.

This stands the set-based methods out for an appealing choice. However, the tremendous searching space without purposeful guidance greatly barriers the learning efficiency and robustness with different modality randomization at the beginning. Therefore, recent works alleviate the learning stability via explicit clustering anchors as queries [8], or conduct post-processing centered on EM-optimized clusters [9]. Subsequently, a contradictory issue arose regarding the balance between the number of modalities with guidance, and the sample efficiency necessary for updating each individual modality.

In this study, we aimed to address these contradictions through the implementation of a group-wise assignment strategy for each modality, referred to as **GTR-R36**. To elaborate further, as depicted in Figure 1, our proposed framework primarily focuses on the decoding stage during predictor training. Given a set of querying modalities as initial inputs, a group-wise transformation is performed, enhancing the input query with corresponding group guidance and facilitating the generation of diverse decoding samples. Subsequently, a shared Transformer decoder stage is

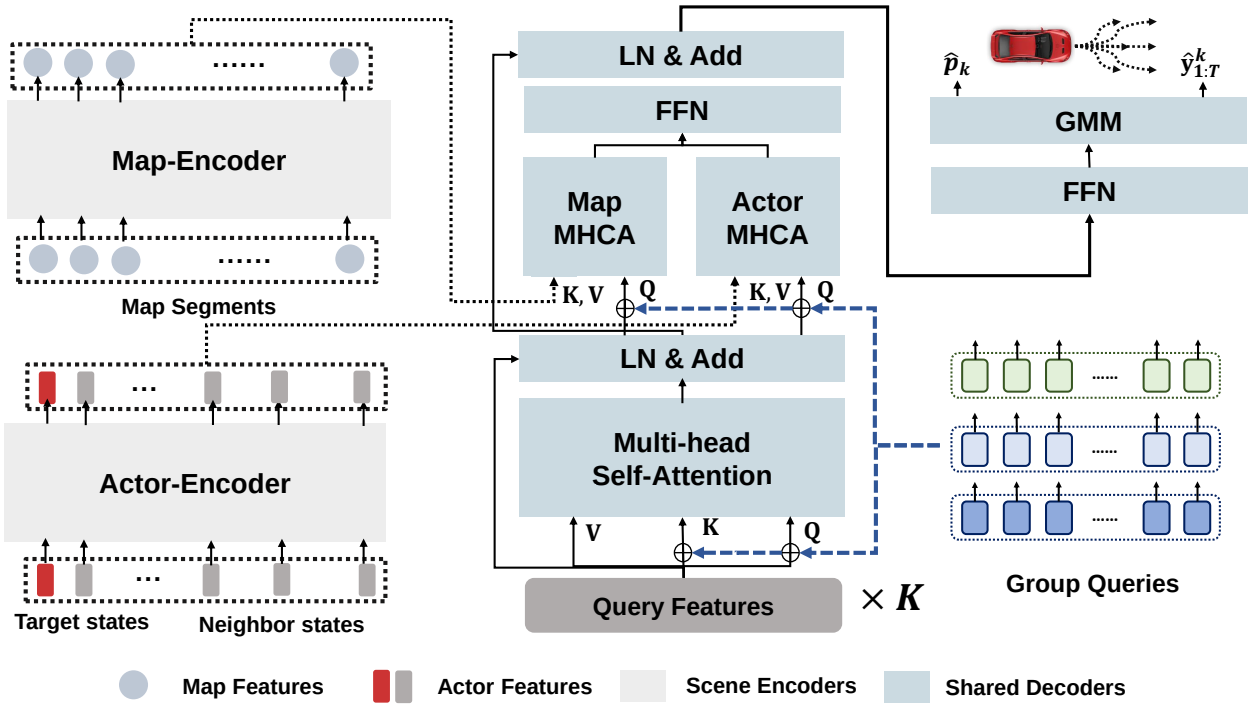

Figure 2. A model framework of the GTR-R36 based on MTR [8] backbone. Scene context inputs are separately encoded Transformer encoders. For iterative decoders, N groups of querying modalities $\mathbf{G}^N$ are respectively initialized and functioned as modality guidance query to be added in query features of Transformer decoder. They conduct shared decoding for final motion predictions founded by Gaussian Mixture-Models (GMM).

employed, enabling the predictor to simultaneously generate motion trajectories across groups, considering the selected modalities. Lastly, a miss-rate optimization process is conducted for each trajectory within the group, promoting differentiation from the hit region within the guidance realm of the selected modality. Thanks to the sample efficiency and differentiation achieved through our proposed GTR approach, the framework exhibits computational cost-efficiency by not requiring additional model parameters and solely utilizing a single group during inference. The proposed paradigm also demonstrates a high level of generality, making it applicable to the majority of current state-of-the-art motion prediction frameworks during the decoding phase.

2. Method

2.1. Scene Context Encoding

The structural overview of proposed GTR following MTR backbone is an encoder-decoder paradigm shown in Figure. 2. With arbitrary scene-level encoding design, we formulate several groups of querying modalities, and conduct shared decoding for group-wise multimodal motion predictions.

The input scene representations mainly encompass two categories: map segments of scene environments and historical actor states encapsulating dynamical information. We adopted the agent-centric pipeline [5] and conduct the transformation centered on target actor to be predicted. Historical states of target actor and N_a closest neighbors are gathered to be encoded in actor features $\mathcal{A} \in R^{N_a \times C}$; Similarly, N_m closest map segments are compressed in a tensor of map features $\mathcal{M} \in R^{N_m \times C}$. We follow [8] and [9] to conduct late fusion during encoding. Separated Transformer encoders are stacked K levels of local feature fusion for $\mathcal{A}$ and $\mathcal{M}$.

2.2. Group-wise Modal Assignment

Given encoded scene context features $\mathcal{A}$ and $\mathcal{M}$, originally we conduct the multimodal prediction by maintaining a set of M modalities as Transformer decoder querying: $\mathbf{Q_U} = \{\mathbf{q}_U^1, \mathbf{q}_U^2, \ldots, \mathbf{q}_U^M\} \in R^{M \times 2}$, where U denotes the type of target actors (vehicle, pedestrian, or cyclist). Each query element of $\mathbf{q}_U^M$ would be responsible to represent certain motion behaviors of the target actor's predictions. The query can be either dynamically initialized, or encoded by clustering of static anchors. We follow MTR to generate each query by k-means clustering: $\mathbf{q}_U^M = \text{SinePos}(I_U^M) \in R^C$, where SinePos is the sinusoidal position encoding operations , and $I_U^M \in R^2$ denotes

the M^{th} clustering for actor type U.

It then adaptive queries the scene features via cross-attention mechanism (MHCA), and utilizes the self-attention module to differentiate each other. However, with growing numbers of modes M which better represent the prediction behaviors, it becomes less of the training samples to update for each modality. Following such motivations, we introduce N groups of queries $\mathbf{G}^N$ so that each modalities can now be updated by N guided samples augmented from the same modality guidance with little differences.

More specifically, the group-wise assignment primarily maintains N groups of modalities, each of which contains all types of modality queries $\mathbf{Q}_U^M$ with respective transformation of MLPs:

$$\begin{aligned} \mathbf{G} =& \{\mathbf{G}^1, \mathbf{G}^2, \ldots, \mathbf{G}^N\} \in R^{N\times M\times C} \\ \mathbf{G}^N =& \mathrm{MLP}^N(\mathbf{Q}_U^M) \end{aligned} \tag{1}$$

In practise, the initialization process of group assignment is done with a linear transformation of $\mathbf{W} \in R^{C\times NC}$ given expanded querying $\mathbf{Q}_U$. Then the group-wise dimension is transposed and reshaped in the batch dimension for shared decoding by iterative Transformer decoders shown in Figure. 2. We modeled the multimodal prediction output using Gaussian Mixture Model (GMM): $\hat{\mathbf{Y}} = \sum_k^M \hat{p}_k \mathcal{N}(\hat{y}_{1:T}^k)$, where $\hat{y}_{1:T}^k$ is the predicted trajectories each for a Bivariate Gaussian elements: $(\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)_{1:T}$. An NLL loss of GMM is devised for each decoder layer and are summed together eventually.

During inference, we can select the final decoder results considering a concatenation of all groups. However, as each modality has already considered the variations across groups during training, in this work we simply drop the rest of groups and only maintain one group of queries ($\mathbf{G}^1$).

2.3. Miss-Rate Optimization

With group-wise assignment of each modality for motion predictions that augment the positive predictions with slight varying by different group transformations, another issue is to delineate the missed predictions from the hit one across groups to boost the trade-offs between predicted precision and recall. Therefore, in our framework we proposed a miss-rate optimization loss for the update across group assigned modality. More specifically, given the predicted states $\hat{s}_t = (\mu_x, \mu_y)_t$ transformed and centered on its corresponding ground-truth s_t as:

$$\hat{s}_t^{\mathbf{T}} = (\hat{s}_t - s_t)\mathbf{R}_t \tag{2}$$

where $\mathbf{R}_t$ is the rotation matrix according to ground-truth headings. The hit region of ground-truth is defined as a bounding box for $\hat{s}_t^T$ under $\mathrm{Thres}^t = (\mathrm{Thres}_x^t, \mathrm{Thres}_y^t)$ scaled by S of current speed v_0:

$$S(v_0) = \mathrm{clip}(0.5 + 0.5\frac{v_0 - 1.4}{11 - 1.4}, 0.5, 1) \tag{3}$$

We utilize a max-margin loss to penalize the transformed predictions that outside the threshold box:

$$\mathbf{c}_t^{\mathrm{mm}} = \begin{cases} |\hat{s}_t^{\mathbf{T}}| - \mathrm{Thres}^t\, S, |\hat{s}_t^{\mathbf{T}}| > \mathrm{Thres}^t\, S, \\ 0,\ \text{otherwise}\,. \end{cases} \tag{4}$$

Given the selected modality, we update the miss-rate loss for all predictions across groups and prediction horizons $T \in \{29, 49, 79\}$: $\mathcal{L}_{mr} = \sum_i^N \sum_t^T \mathbf{c}_t^{\mathrm{mm}\,i}$

2.4. Implementation details

The GTR structure is trained from scratch by WOMD [2] training set, and we consider each targeted actor track to predict as an independent training sample for processing. We choose GELU as the activation function, dropout is added after each layer with a dropout rate of 0.1. We use a distributed training strategy on 4 Tesla A100 with each with a batch size of 96. AdamW optimizer is used with an initial learning rate of 1e-4, and the learning rate decays by a factor of 50% every 2 epochs after 20. The total training epochs are set to 30. Due to limited deadline time, we only validated the GTR-R36 version, namely $N = 3$ of groups and $K = 6$ of decoder layers. We maintain $M = 64$ of modalities for multimodal predictions.

Furthermore, it is worth mentioning that the proposed GTR pipeline is highly versatile and can be easily adapted to other renowned predictors that employ the multimodal prediction paradigm. Therefore, we are planning to validate the generality of proposed approach in future work.

3. Results

Table 1 presents the quantitative outcomes in comparison to other methods on the 2023 Waymo Motion Prediction Leaderboard. The results demonstrate exceptional mean average precision (mAP) scores for the predicted trajectories, leading our method to secure the 3^{rd} position on the leaderboard. It is worth mentioning that there is a concurrent method named GTR in the competition; However, our approach has achieved superior results without relying on any ensemble techniques.

Table 1. Testing Results on 23' Motion Prediction Leaderboard

Method	**Soft mAP**	mAP	minADE	minFDE	MR
MTR++Ens	0.4738	0.4634	0.5581	1.1166	0.1276
IAIR+	0.448	0.4347	0.5783	1.1679	0.1263
Ours	**0.4384**	**0.4255**	**0.6005**	**1.2225**	**0.1279**
GTR	0.4365	0.423	0.5871	1.2096	0.1272
DM	0.4362	0.4301	0.6293	1.2723	0.1473

To further manifest the performance lifting by proposed group-wise pipeline, we conducted an ablation study in examining the group assignment and miss-rate optimization for MTR backbone. The results presented in Table 2

Table 2. Testing results of each prediction type for ablation baselines.

Ablation baseline	Actor Types	**Soft mAP**	mAP	minADE	minFDE	MR
GTR-R36	Vehicle	0.465	0.4521	0.745	1.5049	0.1477
	Pedestrian	0.4357	0.4243	0.347	0.7221	0.0741
	Cyclist	0.4144	0.4003	0.7095	1.4406	0.1772
	Avg	**0.4384**	**0.4255**	**0.6005**	**1.2225**	**0.133**
GTR-3	Vehicle	0.4667	0.4533	0.7415	1.4941	0.1463
	Pedestrian	0.432	0.4205	0.3476	0.7226	0.0754
	Cyclist	0.4087	0.3918	0.7	1.4017	0.1778
	Avg	**0.4358**	**0.4219**	**0.5964**	**1.2061**	**0.1332**
MTR [8]	Vehicle	0.459	0.4494	0.7642	1.5257	0.1514
	Pedestrian	0.4409	0.4331	0.3486	0.727	0.0753
	Cyclist	0.365	0.3561	0.7022	1.4093	0.1786
	Avg	**0.4216**	**0.4129**	**0.605**	**1.2207**	**0.1351**

demonstrate a notable improvement in performance across all metrics when comparing our approach, MTR-3 (assigning 3 groups for simultaneous predictions), to the vanilla MTR method. Notably, the Soft mAP metric shows a notable 3% performance gain, and it is reasonable to anticipate even better results by increasing the number of assigned groups. Additionally, the inclusion of the miss rate optimization loss contributes to an overall precision improvement, resulting in a 1% boost in the final metric. It is important to acknowledge the limitation of having only three groups of varied predictions, as incorporating more predictions would further enhance the effectiveness of the designed miss rate optimization in distinguishing missed predictions.

4. Conclusions

In this study, we introduce a novel group-wise assignment paradigm (named **GTR-R36**) for querying modalities and incorporate miss-rate optimization into multimodal motion predictions. Building upon the MTR backbone, our proposed strategy achieves a substantial 3% improvement using only three groups of assignments, without the need for additional ensemble models. The efficacy of the group assignment approach is further confirmed through rigorous ablation studies. An important aspect of our work is its high generality, as the proposed paradigm can be applied to nearly all types of multimodal predictor structures. We anticipate that further investigation into the group mechanism will lead to even higher results incorporating larger number of groups.

References

[1] Long Chen, Yuchen Li, Chao Huang, Bai Li, Yang Xing, Daxin Tian, Li Li, Zhongxu Hu, Xiaoxiang Na, Zixuan Li, et al. Milestones in autonomous driving and intelligent vehicles: Survey of surveys. *IEEE Transactions on Intelligent Vehicles*, 2022. 1

[2] Scott Ettinger et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021. 3

[3] Thomas Gilles, Stefano Sabatini, Dzmitry Tsishkou, Bogdan Stanciulescu, and Fabien Moutarde. Gohome: Graph-oriented heatmap output for future motion estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 9107–9114. IEEE, 2022. 1

[4] Junru Gu, Chen Sun, and Hang Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15303–15312, 2021. 1

[5] Zhiyu Huang, Haochen Liu, and Chen Lv. Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. *arXiv preprint arXiv:2303.05760*, 2023. 1, 2

[6] Xiaoyu Mo et al. Multi-agent trajectory prediction with heterogeneous edge-enhanced graph attention network. *IEEE Transactions on Intelligent Transportation Systems*, 2022. 1

[7] Sajjad Mozaffari, Omar Y Al-Jarrah, Mehrdad Dianati, Paul Jennings, and Alexandros Mouzakitis. Deep learning-based vehicle behavior prediction for autonomous driving applications: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):33–47, 2020. 1

[8] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *arXiv preprint arXiv:2209.13508*, 2022. 1, 2, 4

[9] 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 *2022 International Conference on Robotics and Automation (ICRA)*, pages 7814–7821. IEEE, 2022. 1, 2