EDA: Evolving and Distinct Anchors for Multimodal Motion Prediction

Longzhong Lin[1,2], Xuewu Lin[2], Tianwei Lin[2], Lichao Huang[2], Rong Xiong[1], Yue Wang[1]*

[1]Zhejiang University
[2]Horizon Robotics
linlongzhong2000@zju.edu.cn, {xuewu.lin, tianwei.lin, lichao.huang}@horizon.cc, {rxiong, ywang24}@zju.edu.cn

Abstract

Motion prediction is a crucial task in autonomous driving, and one of its major challenges lands in the multimodality of future behaviors. Many successful works have utilized mixture models which require identification of positive mixture components, and correspondingly fall into two main lines: prediction-based and anchor-based matching. The prediction clustering phenomenon in prediction-based matching makes it difficult to pick representative trajectories for downstream tasks, while the anchor-based matching suffers from a limited regression capability. In this paper, we introduce a novel paradigm, named Evolving and Distinct Anchors (EDA), to define the positive and negative components for multimodal motion prediction based on mixture models. We enable anchors to evolve and redistribute themselves under specific scenes for an enlarged regression capacity. Furthermore, we select distinct anchors before matching them with the ground truth, which results in impressive scoring performance. Our approach enhances all metrics compared to the baseline MTR, particularly with a notable relative reduction of 13.5% in Miss Rate, resulting in state-of-the-art performance on the Waymo Open Motion Dataset. Appendix and code are available at https://github.com/Longzhong-Lin/EDA.

Introduction

In the field of autonomous driving, motion prediction is an important task which contributes to scene understanding and safe planning. Motion prediction utilizes historical agent states and road maps to predict the future trajectories of traffic participants. In recent years, an increasing amount of research works (2023; 2023; 2022a; 2022; 2021; 2021; 2020; 2019; 2018; 2017) have focused on motion prediction. A major challenge of motion forecasting is the multimodality of future behaviors, which means an agent could carry out one of many underlying possibilities.

A bunch of works (Ngiam et al. 2021; Varadarajan et al. 2022; Shi et al. 2022a; Chai et al. 2019) have adopted mixture models, like Gaussian Mixture Model (GMM), to represent multimodal future behaviors and have gained great success, where potential trajectories are modeled as scored components. These approaches typically employ a winner-takes-all regression loss in conjunction with a classification term, which necessitates identifying the positive and negative mixture components. For selecting positive components, there are two main categories of existing methods: **prediction-based** and **anchor-based** matching.

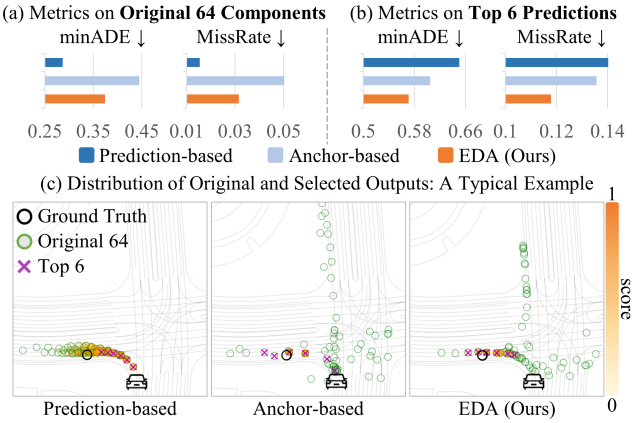


Figure 1: The outcomes from different matching paradigms. All of the strategies share the same network structure with 64 learnable queries. The top 6 predictions are selected from the original ones by non-maximum suppression (NMS).

The prediction-based matching methods (Ngiam et al. 2021; Varadarajan et al. 2022) choose the predicted trajectory that is closest to the ground truth as the positive component, which is demonstrated in Fig. 2(a). Predictions generated by these methods honestly reflect the high degree of uncertainty in future behaviors, which results in an originally lower minimum error and miss rate (Fig. 1(a)). However, as illustrated in Fig. 1(c), the output trajectories from prediction-based matching tend to cluster around the most probable regions and similar scores are made upon such predictions, making it difficult to pick representative trajectories for downstream tasks (Fig. 1(b)).

As demonstrated in Fig. 2(b), the anchor-based matching methods (Shi et al. 2022a; Chai et al. 2019) associate each component with an anchor endpoint or trajectory, and select the positive one corresponding to the closest predefined anchor to ground truth. The introduction of spatial priors considerably alleviates the burden of optimization in classification, and the methods would prefer to generate trajecto-

*Corresponding author.

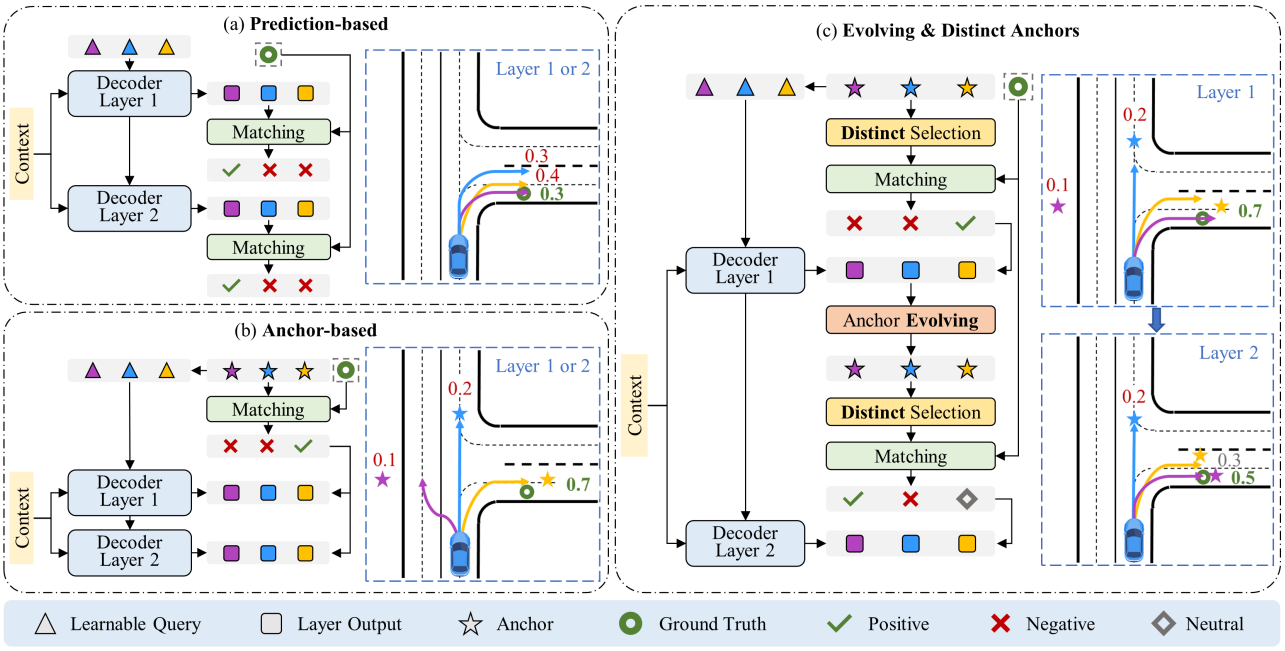


Figure 2: The demonstration of different matching paradigms with a 2-layer decoder. Each subfigure displays a workflow on the left and corresponding illustration on the right. Objects with the same internal color belong to the same mixture component. The numbers attached to each component represent the scores. (a) and (b) respectively present the *prediction-based* and *anchor-based* matching. (c) demonstrates the design of proposed *Evolving and Distinct Anchors (EDA)*, where the anchors for the 2nd layer are updated using the outputs from the 1st layer. Additionally, a selection of distinct anchors is applied before matching. As a result, the yellow component in the 2nd layer is excluded since it is close to the purple one but has a lower score.

ries around the predefined anchors. Nevertheless, to reduce computational costs and prevent compromising the scoring performance (Shi et al. 2022a), the anchors are usually distributed in a sparser manner compared to the outputs from prediction-based matching. Hence the regression capability of model is limited, which is shown in Fig. 1(a).

In this paper, we introduce a novel paradigm, named **Evolving and Distinct Anchors (EDA)**, to define the positive and negative components for multi-modal motion prediction based on mixture models. As illustrated in Fig. 2(c), we first pre-define anchors and then update them by the intermediate outputs, hence the name **Evolving Anchors**. On the one hand, we utilize spatial priors in the form of predefined anchors to alleviate the difficulties in trajectory scoring posed by prediction-based matching approaches. On the other hand, we allow anchors to redistribute themselves based on predictions under specific scenes for a promoted regression capability compared to the vanilla anchor-based matching. As the anchors evolve multiple times, we observe that the prediction clustering issue previously presented in prediction-based matching arises and becomes pronounced, which continues to bother the optimization in scoring trajectories. In order to mitigate the ambiguity in classification caused by the gathering problem, inspired by Dense Distinct Query (Zhang et al. 2023) for object detection, we select **Distinct Anchors** through non-maximum suppression (NMS) before matching them with the ground truth, as demonstrated in Fig. 2(c). The adoption of distinct anchors also encourages the model to prioritize the most probable component among similar ones, facilitating the selection of representative predictions for downstream jobs. It turns out that our method leverages the benefits of both anchor-based and prediction-based matching (as shown in Fig. 1), and achieves state-of-the-art performance on the Waymo Open Motion Dataset (Ettinger et al. 2021).

Our contributions can be summarized as follows:

1. We propose the Evolving Anchors for multimodal motion prediction based on mixture models, where we predefine spatial anchors and then update them by the intermediate outputs. This novel strategy strikes a balance between the existing anchor-based and prediction-based matching approaches.
2. We adopt Distinct Anchors to address the ambiguity in classification induced by prediction clustering phenomena. Employing NMS on anchors before matching them with the ground truth, we reduce the optimization difficulty in trajectory scoring and enhance the selection of representative predictions for subsequent tasks.
3. We have performed experiments on the Waymo Open Motion Dataset (2021). With the assistance of Evolving and Distinct Anchors, our single model has surpassed the performance of previous ensemble-free approaches, exhibiting improvements on all metrics compared to the baseline MTR (Shi et al. 2022a), particularly with a significant relative reduction of 13.5% in Miss Rate.

Related Work

Architectures for Motion Prediction

In recent times, there has been a significant increase in the study of motion prediction owing to the rising interest in autonomous driving. Motion prediction involves using the past agent states and road maps to forecast the future paths of traffic participants. Early studies (Chai et al. 2019; Casas et al. 2020; Park et al. 2020; Gilles et al. 2021; Casas, Sadat, and Urtasun 2021) commonly rasterize the inputs into images and capture the contextual information through CNNs. LaneGCN (Liang et al. 2020) and LaneRCNN (Zeng et al. 2021) construct lane graphs to efficiently represent the topology of road maps. Recent works (Gu, Sun, and Zhao 2021; Varadarajan et al. 2022; Shi et al. 2022a) have widely adopted the VectorNet (Gao et al. 2020) representation scheme, which regards the road maps as polylines. As Transformers (Vaswani et al. 2017) have gained popularity, an increasing number of studies (Liu et al. 2021; Ngiam et al. 2021; Jia et al. 2023) have utilized the attention mechanism to encode scene context. Encouraged by the successful application of DETR (Carion et al. 2020), many Transformer-based models (Girgis et al. 2021; Varadarajan et al. 2022; Nayakanti et al. 2023) have adopted learnable queries in decoder to generate multiple potential future trajectories. In our study, we utilize the architecture presented in MTR (Shi et al. 2022a), which is an advanced transformer framework incorporating a local attention based encoder and a decoder with intention queries.

Modeling for Multimodal Future Motion

Previous studies have investigated different approaches for modeling multimodal future behaviors. Earlier generative models (Lee et al. 2017; Gupta et al. 2018; Rhinehart, Kitani, and Vernaza 2018; Rhinehart et al. 2019) generate a collection of samples to represent the distribution of future. Many other works (Chai et al. 2019; Mercat et al. 2020; Ngiam et al. 2021) have utilized mixture models to parameterize multi-modal predictions, which mainly fall into two lines: prediction-based and anchor-based matching, as elaborated in introduction. In prediction-based matching methods (Ngiam et al. 2021; Varadarajan et al. 2022; Nayakanti et al. 2023), the positive mixture component is chosen by directly comparing predicted trajectories to the ground truth. Some models (Tang and Salakhutdinov 2019; Girgis et al. 2021) using the loss based on EM algorithm can also be viewed as prediction-based matching when its KL term converges. Due to the challenge of selecting representative future trajectories, these methods have opted to use well-designed aggregation techniques (Varadarajan et al. 2022; Nayakanti et al. 2023), or to directly utilize an end-to-end version (Ngiam et al. 2021; Girgis et al. 2021). However, their scoring performance still lags behind that of anchor-based matching methods. The anchor-based matching (Chai et al. 2019; Zhao et al. 2021) regards as positive the component matching the closest predefined anchor to ground truth. The HOME series (Gilles et al. 2021, 2022) and DenseTNT (Gu, Sun, and Zhao 2021) can be considered as variations of anchor-based matching, where the anchors are the grids in heatmaps or target candidates placed on roads, but they require an additional sampling process to obtain the final predictions. The MTR (Shi et al. 2022a) achieves remarkable scoring performance using predefined anchors, while its end-to-end prediction-based matching version demonstrates significantly better performance in terms of minimum error and miss rate. Motivated by the findings, we propose a novel matching paradigm to exploit the regression potential hidden by the state-of-the-art anchor-based matching strategy.

Dense Distinct Query for Label Assignment

According to Zhang et al., considering one-to-one label assignment in object detection, sparse queries cannot ensure a high recall, while dense queries inevitably bring more similar queries and face optimization challenges in classification. Therefore, they propose Dense Distinct Queries (DDQ), in which dense queries are first laid and then distinct queries are selected for one-to-one assignments. Inspired by DDQ (Zhang et al. 2023), we adopt distinct anchors to mitigate the ambiguity in trajectory scoring induced by prediction clustering phenomena.

Evolving and Distinct Anchors

For identifying positive components, there are two primary strategies within the existing mixture-model based methods. The *prediction-based* matching directly compares the predicted trajectories $\{P_i\}_{i=1}^{N_C}$ with the ground truth G:

$$Distance(P_i, G),\ i = 1, \cdots, N_C, \tag{1}$$

where N_C denotes the number of components. In *anchor-based* matching, the spatial anchors $\{A_i\}_{i=1}^{N_C}$ are linked to each component and matched with the ground truth G:

$$Distance(A_i, G),\ i = 1, \cdots, N_C. \tag{2}$$

In this study, we present *Evolving and Distinct Anchors (EDA)*, a novel paradigm to define the positive and negative mixture components by:

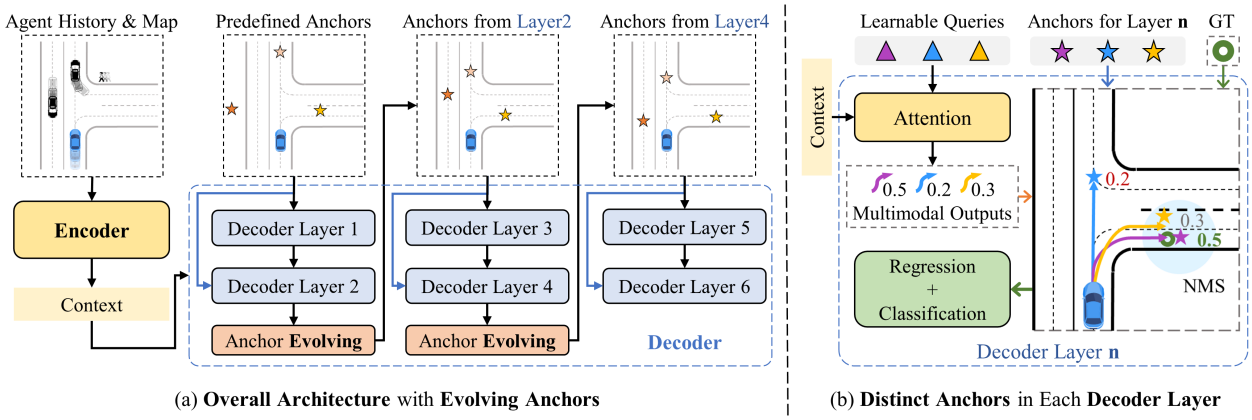
$$Distance(A_{E_j}, G),\ j \in \mathcal{I}_D, \tag{3}$$

where A_E denotes the evolving anchors, and $\mathcal{I}_D$ is the index set of distinct anchors. The main idea is illustrated in Fig. 3. In the following we first introduce the encoder-decoder structure upon which our method is built. Subsequently, we provide detailed descriptions of the proposed *Evolving Anchors* and *Distinct Anchors* respectively.

Network Architecture

We have implemented our ideas on a cutting-edge encoder-decoder structure, as the one presented in MTR (Shi et al. 2022a). This transformer framework employs an encoder with local self-attention for scene context modeling, in addition to a multi-layer decoder that incorporates learnable intention queries to predict multimodal trajectories.

It is important to note that our approach presented in this paper is centered on the design of loss. Consequently, the proposed *Evolving and Distinct Anchors (EDA)* can be readily applied to any network structure that includes a multi-layer decoder.



(a) **Overall Architecture** with **Evolving Anchors** (b) **Distinct Anchors** in Each **Decoder Layer**

Figure 3: The illustration of the EDA paradigm. (a) shows an instance of the overall architecture with a 6-layer decoder and anchors evolving at the 2nd, 4th layers. (b) reveals the details in each decoder layer, where distinct anchors are selected before matching. Components that correspond to the excluded anchors, such as the yellow one in picture, are considered neutral.

Evolving Anchors

Although the spatial priors significantly alleviate the challenge in classification optimization, the vanilla anchor-based matching encounters a limitation in its regression capability, which will be demonstrated later. Regarding the above issue and encouraged by the successful adoption of multi-layer decoders in motion prediction (2021; 2022a), we naturally consider enabling anchors to evolve through multiple decoder layers for an enlarged regression capacity.

Take a 6-layer decoder for instance, as illustrated in Fig. 3(a), we can implement twice-evolving anchors by updating the anchors with outputs from the 2nd and 4th layers, in which the evolving anchors for the n-th layer are:

$$A_E^{(n)} = \begin{cases} A, & n = 1, 2 \\ P^{(2)}, & n = 3, 4 \\ P^{(4)}, & n = 5, 6 \end{cases}, \tag{4}$$

where we have omitted the index subscripts for simplicity.

In a word, the evolving anchors are *initially predefined* and *later adjusted* by the intermediate outputs from decoder layers, which means the anchors are allowed to redistribute themselves under specific scenes.

Effects of Evolving Anchors. The vanilla anchor-based matching, as presented in Fig. 4, tends to make relative small adjustments to the predefined anchors in each layer. This is because, making significant changes to the anchor that hits the ground truth would result in a considerable regression loss, while the refinements to unlikely ones are not encouraged. Besides, the anchors are usually distributed in a sparser manner to reduce computational costs and avoid compromising the scoring performance (Shi et al. 2022a). Therefore, the regression capability of model is limited by the anchor-based matching with static anchors.

Correspondingly, making anchors adjustable motivates the model to modify unreasonable components in a larger degree, as illustrated in Fig. 4. Nevertheless, substantial refinements are made only when the potential benefits of

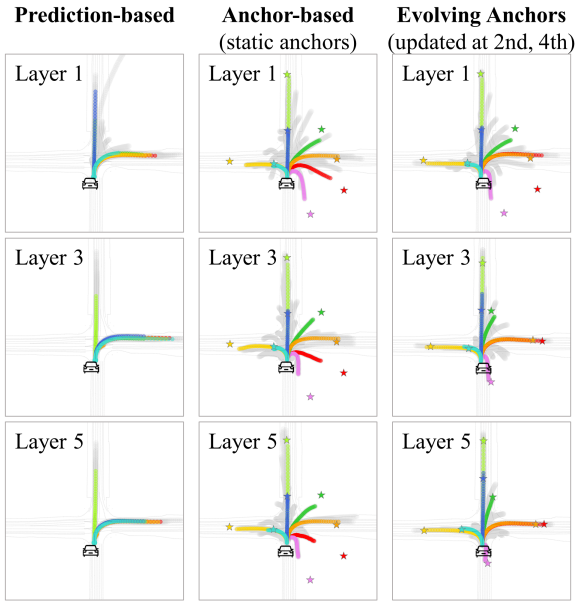


Figure 4: Layer outputs from different matching paradigms under the same scene. The ⋆ represents the anchor endpoint. The typical trajectories are highlighted in bright colors, with each color indicating the same component across various methods, whereas the remaining ones are displayed in gray.

achieving successful regression outweigh the expected cost of mistakenly making substantial adjustments. Hence the modifications to anchors are restrained and progressive in evolving anchors. In contrast, without the constraints from predefined anchors, the prediction-based matching would generate trajectories gathering around the most possible regions, even in the earlier layers, as shown in Fig. 4.

Therefore, the proposed *Evolving Anchors* achieves a balance between the anchor-based and prediction-based matching, where one can adjust the extent of modifications to predefined anchors through the frequency of anchor updates.

Distinct Anchors

Although predicting trajectories that cluster around the most probable regions contributes to better coverage of future behaviors with high uncertainty in prediction-based matching, this preference also introduces a serious issue of ambiguity in the scoring task. With multiple gathering outcomes, it becomes difficult for the model to distinguish the actual one closest to the ground truth. Hence the model tends to output similar scores for such predictions, making it hard to pick representative trajectories for downstream tasks.

In our proposed evolving anchors, as stated in the above analysis on *effects of evolving anchors*, the more frequently we update anchors, the greater the opportunity for substantial adjustments to unreal components. However, this also increases the potential for the phenomenon of prediction clustering. Such patterns can be observed intuitively in Fig. 5. As a result, this issue continues to pose a challenge for optimization in classification, particularly when updating the anchors multiple times.

Taking inspiration from DDQ (Zhang et al. 2023) in the object detection domain, we attempt to adopt distinct anchors to improve scoring performance. Specifically, we apply non-maximum suppression (NMS) to the anchors for each decoder layer prior to matching them with the ground truth during training, as illustrated in Fig. 3(b). Mixture components that correspond to the excluded anchors will neither serve as positive nor negative samples. Through the aforementioned operations:

- We prevent the labeling of similar anchors as opposite, which significantly reduces the optimization difficulty for the classification task.
- Moreover, the model is encouraged to prioritize the most probable trajectory among the similar ones, making it easier to select the representative predictions using simple post-processing techniques such as NMS.

Training Losses

We train the model with a combination of winner-takes-all regression loss and classification term, which is commonly used in mixture-model based methods (Chai et al. 2019; Nayakanti et al. 2023). Same as MTR (Shi et al. 2022a), we employ a Gaussian regression loss. Instead of Cross Entropy (CE) in MTR, we use Binary Cross Entropy (BCE) for classification loss, which is suitable for arbitrary numbers of mixture components filtered by distinct anchors. Please refer to the Appendix (Lin et al. 2023) for more implementation details.

Experiments

Experimental Setup

Dataset and metrics. We assess our method on the large-scale Waymo Open Motion Dataset (WOMD) proposed by Ettinger et al., which extracts interesting behaviors from actual traffic scenes. The WOMD (Ettinger et al. 2021) includes 487k training scenes, 44k validation and 44k testing scenes, where each scene contains up to 8 target agents. Each agent is comprised of 1 second of historical states and 8 seconds of future information. The long time horizon challenges the model's capacity to capture a broad field of view and adapt to a vast output space for trajectories.

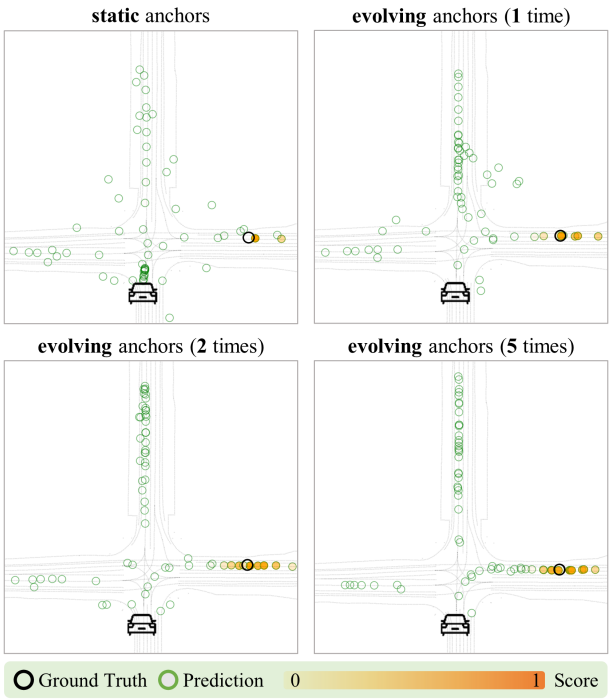


Figure 5: A typical example illustrating the prediction clustering phenomenon in evolving anchors.

Due to the complexity of reasoning about numerous potential future behaviors, benchmark metrics limit the number of trajectories under consideration. The official website offers an evaluation on submissions with up to 6 motion predictions for each target agent, returning metrics including minADE (Minimum Average Displacement Error), minFDE (Minimum Final Displacement Error), Miss Rate, Overlap Rate, mAP and Soft mAP. Hence the top 6 metrics we provide are obtained from the official evaluation server, whereas we utilize a local evaluation tool based on the official API to compute metrics on a greater number of mixture components.

Implementation details. Our design is built upon the state-of-the-art MTR framework (Shi et al. 2022a), where we adopt the default setting of the network structure and training configuration. We train the model for 30 epochs on 16 GPUs (NVDIA RTX 3090) with the batch size of 80 scenes. The predefined anchors we use are the 64 intention points generated by a k-means clustering algorithm on the training set, as used in MTR. To achieve a more stable matching, except for predefined anchors we assign labels based on the full trajectories of intermediate outputs that act as evolving anchors.

For evaluation, we pick top 6 predictions by employing NMS on the endpoints of 64 predicted trajectories. Following Shi et al., the distance threshold σ is scaled proportion-

Anchor Evolving Times	Classification Loss	Distinct Anchors	mAP ↑			minADE ↓	minFDE ↓	Miss Rate ↓
			original	scaled	rank			
0	CE		0.4059	0.4167	0.4121	0.6012	1.2277	0.1348
0	BCE		0.4053	0.4171	0.4126	0.6050	1.2376	0.1357
1	CE		0.4013	0.4211	0.4183	0.5867	1.2109	0.1240
1	BCE		0.4060	**0.4255**	0.4228	0.5838	1.2012	0.1221
1	BCE	✓	0.4173	0.4221	0.4278	0.5776	1.1895	0.1203
2	CE		0.3868	0.4107	0.4101	0.5881	1.2145	0.1227
2	BCE		0.3957	0.4236	0.4207	0.5888	1.2144	0.1229
2	BCE	✓	**0.4235**	0.4251	**0.4353**	**0.5708**	**1.1730**	**0.1178**
5	CE		0.3647	0.4051	0.4002	0.5996	1.2444	0.1264
5	BCE		0.3675	0.4063	0.4037	0.5998	1.2412	0.1272
5	BCE	✓	0.4186	0.4185	0.4322	0.5817	1.2056	0.1245

Table 1: Top 6 metrics on the validation set of Waymo Open Motion Dataset (Ettinger et al. 2021). The terms "original", "scaled" and "rank" under the "mAP" heading respectively represent the results upon the original, scaled and ranking-oriented top 6 scores, as elaborated in implementation details.

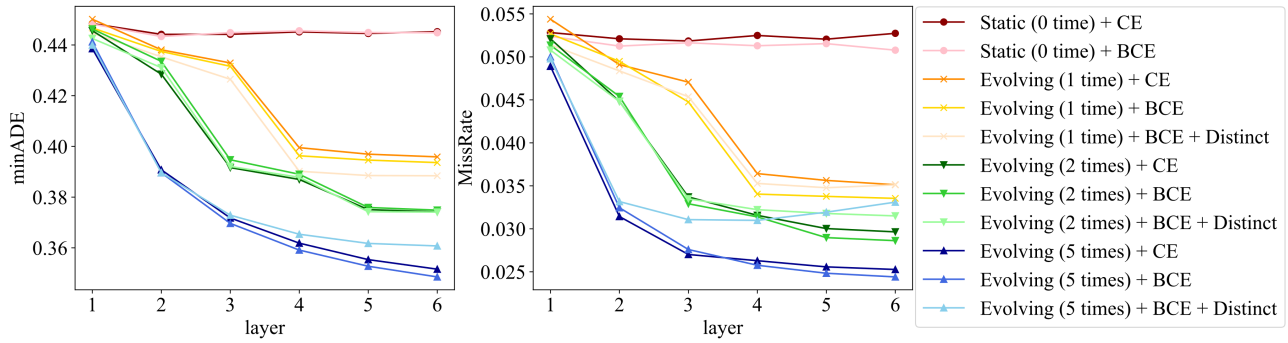


Figure 6: Minimum Error (left) and Miss Rate (right) on original 64 components for each decoder layer.

ally to the length L of trajectory with the highest confidence: $\sigma = \min[3.5, \max[2.5, 2.5 + 1.5 \times (L - 10)/(50 - 10)]]$. The same NMS distance threshold is also applied to the selection of distinct anchors. To improve the mAP metrics, the MTR (2022a) scales the original top 6 scores for each sample through dividing them by their sum, making the scores comparable across different agents. As far as we are concerned, it also makes sense to consider the rank of trajectories in a sample when comparing predictions across different agents. Therefore, We add a rank-related integer to the original scores ranging between 0 and 1, to ensure that when computing the mAP metrics, the top-ranked trajectories of all samples will be sorted at the top, followed by the 2nd-ranked, 3rd-ranked, and so on. For instance, we add 5 for the top-ranked trajectory, 4 for the 2nd-ranked, 3 for the 3rd-ranked, and so forth. In order to align with previous works, we still present the mAP metrics upon the original and scaled scores in the following ablation study.

Ablation Study

We first investigate the impacts of *Evolving Anchors*, and then assess the effectiveness of *Distinct Anchors*. All models are evaluated on the validation set of WOMD (Ettinger et al. 2021). In terms of mAP metrics, the results based on the original, scaled, and ranking-oriented top 6 scores are all presented, as referred in implementation details.

Evolving Anchors. Starting from the baseline with 0 time of anchor updating, which is actually the MTR (Shi et al. 2022a) that uses the anchor-based matching with static anchors, we apply various *anchor evolving times* to explore the effects of evolving anchors. Upon the adopted 6-layer decoder, we update anchors at the 3rd layer for once-updating anchors, at the 2nd and 4th layers for twice-evolving anchors, and at every but the final layer for 5 times of anchor evolving. The corresponding top 6 metrics are displayed in the rows highlighted in gray of Table 1, while the results on original 64 components are included in Fig. 6.

Fig. 6 shows that the regression capacity of model improves as the number of anchor updates increases, with a significant enhancement each time the anchors evolve. This finding supports the idea that evolving anchors present opportunities to unlock the potential in regression hidden by the vanilla anchor-based matching. And the more frequently we update the anchors, the greater the potential for adjustments to enhance the regression.

Set	Method	Soft mAP ↑	mAP ↑	minADE ↓	minFDE ↓	Miss Rate ↓	Overlap Rate ↓
Test	MotionCNN (2022)	-	0.2136	0.7400	1.4936	0.2091	0.1560
	ReCoAt (2022)	-	0.2711	0.7703	1.6668	0.2437	0.1642
	DenseTNT (2021)	-	0.3281	1.0387	1.5514	0.1573	0.1779
	SceneTransformer (2021)	-	0.2788	0.6117	1.2116	0.1564	0.1473
	HDGT (2023)	-	0.2854	0.5933	1.2055	0.1511	-
	MTR (2022a)	0.4216	0.4129	0.6050	1.2207	0.1351	0.1277
	MTR++ (2023)	0.4414	0.4329	0.5906	1.1939	0.1298	0.1281
	EDA (Ours)	**0.4510**	**0.4401**	**0.5718**	**1.1702**	**0.1169**	**0.1266**
Val	MTR (2022a)	-	0.4164	0.6046	1.2251	0.1366	-
	MTR++ (2023)	-	0.4351	0.5912	1.1986	0.1296	-
	EDA (Ours)	**0.4462**	**0.4353**	**0.5708**	**1.1730**	**0.1178**	**0.1273**

Table 2: Performance comparison on the validation and test sets of Waymo Open Motion Dataset (Ettinger et al. 2021).

However, as illustrated in Fig. 5, the phenomenon of prediction clustering also becomes severe when the anchors are updated more times, since the increased freedom in modifying the predefined anchors results in outputs more resembling those from the prediction-based matching. This issue adversely affects the performance of trajectory scoring, leading to a decline in top 6 metrics when two or more anchor updates are employed, as presented in Table 1.

Distinct Anchors. We utilize the BCE loss to accommodate varying numbers of the mixture components selected for distinct anchors, which is different from the MTR (Shi et al. 2022a) using the CE loss. Hence we begin by assessing the influence of various options for the classification loss. From both Fig. 6 and Table 1, it can be observed that, overall, the BCE loss leads to only marginal differences in the results, along with a slightly better mAP. This suggests that the BCE loss can be considered a reasonable substitute for the CE classification loss.

After validating the impact of BCE loss, we now evaluate the efficacy of *Distinct Anchors*. As seen in Table 1, the use of distinct anchors brings a considerable enhancement in the top 6 metrics for models with evolving anchors. What's more, the progress, particularly in mAP (*e.g.*, +0.5%, +1.46%, +2.85% for 1, 2, 5 anchor updates respectively upon ranking-oriented scores), becomes notable with a higher frequency of anchor evolving. Nevertheless, the regression metrics on original 64 mixture components, as shown in Fig. 6, do not exhibit a significant improvement. Such evidences indicate that the adoption of distinct anchors does facilitate the selection for the representative behaviors as well as the scoring performance, which is hindered by the prediction clustering phenomenon.

But the benefits of distinct anchors are not limitless. As depicted in Table 1, both with the help of distinct anchors, the performance of 5 anchor updates cannot surpass that of twice-evolving anchors at all. And the unusual deterioration in Miss Rate when using distinct anchors for 5 times of anchor evolving (Fig. 6) implies that the model may be still plagued by too many anchor updates.

Benchmark Results

We evaluate the model that performs the best in our ablation study, namely *twice-evolving and distinct anchors* with ranking-oriented top 6 scores, on the test set of WOMD (Ettinger et al. 2021). We need to point out that the model for testing is trained solely on the WOMD training set without any ensemble techniques applied, consistent with our baseline MTR (Shi et al. 2022a).

As shown in Table 2, our single model outperforms previous ensemble-free approaches on the WOMD. The proposed EDA has demonstrated significant improvements in all performance metrics compared to the baseline MTR on both the validation and test sets. Specifically, there is a relative improvement of 13.5% on Miss Rate, 5.5% on minADE, and 4.1% on minFDE, as well as a +2.94% absolute growth in SoftmAP on the test set. Furthermroe, the performance of our EDA surpasses that of MTR++ (Shi et al. 2023), the latest improved version of MTR, on both the validation and test sets of WOMD. It is worth noting that MTR++ primarily enhances the network structure of MTR, while our approach centers on the design of loss, which means that combining the two complementary refinements has the potential to yield even more remarkable performance. Please refer to the Appendix (Lin et al. 2023) for more experimental results.

Conclusions

In this paper, we present Evolving and Distinct Anchors (EDA), a novel paradigm to define the positive and negative components for multi-modal motion prediction based on mixture models. We pre-define anchors and update them with intermediate outputs, and pick distinct anchors before matching them with the ground truth. Allowing the anchors to evolve and redistribute themselves under specific scenes promotes the regression capability of model. The adoption of distinct anchors addresses the ambiguity in classification induced by the prediction clustering issue, and facilitates the selection of representative predictions for downstream tasks. It turns out that our approach exhibits a significant improvement compared to the baseline MTR, achieving state-of-the-art performance on the Waymo Open Motion Dataset.

Acknowledgements

This work was supported by the National Nature Science Foundation of China under Grant 62373322, in part by the Key R&D Program of Zhejiang (2022C01022, 2023C01176), and in part by the Zhejiang Provincial Natural Science Foundation of China (LD22E050007).

References

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *European conference on computer vision*, 213–229. Springer.

Casas, S.; Gulino, C.; Liao, R.; and Urtasun, R. 2020. Spagnn: Spatially-aware graph neural networks for relational behavior forecasting from sensor data. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 9491–9497. IEEE.

Casas, S.; Sadat, A.; and Urtasun, R. 2021. Mp3: A unified model to map, perceive, predict and plan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14403–14412.

Chai, Y.; Sapp, B.; Bansal, M.; and Anguelov, D. 2019. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction. *arXiv preprint arXiv:1910.05449*.

Ettinger, S.; Cheng, S.; Caine, B.; Liu, C.; Zhao, H.; Pradhan, S.; Chai, Y.; Sapp, B.; Qi, C. R.; Zhou, Y.; et al. 2021. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9710–9719.

Gao, J.; Sun, C.; Zhao, H.; Shen, Y.; Anguelov, D.; Li, C.; and Schmid, C. 2020. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11525–11533.

Gilles, T.; Sabatini, S.; Tsishkou, D.; Stanciulescu, B.; and Moutarde, F. 2021. Home: Heatmap output for future motion estimation. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 500–507. IEEE.

Gilles, T.; Sabatini, S.; Tsishkou, D.; Stanciulescu, B.; and Moutarde, F. 2022. Gohome: Graph-oriented heatmap output for future motion estimation. In *2022 international conference on robotics and automation (ICRA)*, 9107–9114. IEEE.

Girgis, R.; Golemo, F.; Codevilla, F.; Weiss, M.; D'Souza, J. A.; Kahou, S. E.; Heide, F.; and Pal, C. 2021. Latent variable sequential set transformers for joint multi-agent motion prediction. *arXiv preprint arXiv:2104.00563*.

Gu, J.; Sun, C.; and Zhao, H. 2021. Densetnt: End-to-end trajectory prediction from dense goal sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15303–15312.

Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; and Alahi, A. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2255–2264.

Huang, Z.; Mo, X.; and Lv, C. 2022. ReCoAt: A deep learning-based framework for multi-modal motion prediction in autonomous driving application. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, 988–993. IEEE.

Jia, X.; Wu, P.; Chen, L.; Liu, Y.; Li, H.; and Yan, J. 2023. Hdgt: Heterogeneous driving graph transformer for multi-agent trajectory prediction via scene encoding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Konev, S.; Brodt, K.; and Sanakoyeu, A. 2022. MotionCNN: a strong baseline for motion prediction in autonomous driving. *arXiv preprint arXiv:2206.02163*.

Lee, N.; Choi, W.; Vernaza, P.; Choy, C. B.; Torr, P. H.; and Chandraker, M. 2017. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 336–345.

Liang, M.; Yang, B.; Hu, R.; Chen, Y.; Liao, R.; Feng, S.; and Urtasun, R. 2020. Learning lane graph representations for motion forecasting. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, 541–556. Springer.

Lin, L.; Lin, X.; Lin, T.; Huang, L.; Xiong, R.; and Wang, Y. 2023. EDA: Evolving and Distinct Anchors for Multimodal Motion Prediction. arXiv:2312.09501.

Liu, Y.; Zhang, J.; Fang, L.; Jiang, Q.; and Zhou, B. 2021. Multimodal motion prediction with stacked transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7577–7586.

Mercat, J.; Gilles, T.; El Zoghby, N.; Sandou, G.; Beauvois, D.; and Gil, G. P. 2020. Multi-head attention for multi-modal joint vehicle motion forecasting. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 9638–9644. IEEE.

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

Ngiam, J.; Vasudevan, V.; Caine, B.; Zhang, Z.; Chiang, H.-T. L.; Ling, J.; Roelofs, R.; Bewley, A.; Liu, C.; Venugopal, A.; et al. 2021. Scene transformer: A unified architecture for predicting future trajectories of multiple agents. In *International Conference on Learning Representations*.

Park, S. H.; Lee, G.; Seo, J.; Bhat, M.; Kang, M.; Francis, J.; Jadhav, A.; Liang, P. P.; and Morency, L.-P. 2020. Diverse and admissible trajectory forecasting through multimodal context understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, 282–298. Springer.

Rhinehart, N.; Kitani, K. M.; and Vernaza, P. 2018. R2p2: A reparameterized pushforward policy for diverse, precise generative path forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 772–788.

Rhinehart, N.; McAllister, R.; Kitani, K.; and Levine, S. 2019. Precog: Prediction conditioned on goals in visual

multi-agent settings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2821–2830.

Shi, S.; Jiang, L.; Dai, D.; and Schiele, B. 2022a. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35: 6531–6543.

Shi, S.; Jiang, L.; Dai, D.; and Schiele, B. 2022b. MTR-A: 1st Place Solution for 2022 Waymo Open Dataset Challenge – Motion Prediction. arXiv:2209.10033.

Shi, S.; Jiang, L.; Dai, D.; and Schiele, B. 2023. MTR++: Multi-Agent Motion Prediction with Symmetric Scene Modeling and Guided Intention Querying. *arXiv preprint arXiv:2306.17770*.

Tang, C.; and Salakhutdinov, R. R. 2019. Multiple futures prediction. *Advances in neural information processing systems*, 32.

Varadarajan, B.; Hefny, A.; Srivastava, A.; Refaat, K. S.; Nayakanti, N.; Cornman, A.; Chen, K.; Douillard, B.; Lam, C. P.; Anguelov, D.; et al. 2022. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, 7814–7821. IEEE.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ye, M.; Cao, T.; and Chen, Q. 2021. Tpcn: Temporal point cloud networks for motion forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11318–11327.

Zeng, W.; Liang, M.; Liao, R.; and Urtasun, R. 2021. Lanercnn: Distributed representations for graph-centric motion forecasting. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 532–539. IEEE.

Zhang, S.; Wang, X.; Wang, J.; Pang, J.; Lyu, C.; Zhang, W.; Luo, P.; and Chen, K. 2023. Dense Distinct Query for End-to-End Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7329–7338.

Zhao, H.; Gao, J.; Lan, T.; Sun, C.; Sapp, B.; Varadarajan, B.; Shen, Y.; Shen, Y.; Chai, Y.; Schmid, C.; et al. 2021. Tnt: Target-driven trajectory prediction. In *Conference on Robot Learning*, 895–904. PMLR.

Zhou, Z.; Ye, L.; Wang, J.; Wu, K.; and Lu, K. 2022. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8823–8833.