

# AMEND: A Mixture of Experts Framework for Long-tailed Trajectory Prediction

Ray Coden Mercurius<sup>1,3</sup>, Ehsan Ahmadi<sup>2,3</sup>, Soheil Mohamad Alizadeh Shabestary<sup>3</sup>, Amir Rasouli<sup>3</sup>

**Abstract**—Accurate prediction of pedestrians’ future motions is critical for intelligent driving systems. Developing models for this task requires rich datasets containing diverse sets of samples. However, the existing naturalistic trajectory prediction datasets are generally imbalanced in favor of simpler samples and lack challenging scenarios. Such a long-tail effect causes prediction models to underperform on the tail portion of the data distribution containing safety-critical scenarios. Previous methods tackle the long-tail problem using methods such as contrastive learning and class-conditioned hypernetworks. These approaches, however, are not modular and cannot be applied to many machine learning architectures. In this work, we propose a modular model-agnostic framework for trajectory prediction that leverages a specialized mixture of experts. In our approach, each expert is trained with a specialized skill with respect to a particular part of the data. To produce predictions, we utilise a router network that selects the best expert by generating relative confidence scores. We conduct experimentation on common pedestrian trajectory prediction datasets and show that besides achieving state-of-the-art performance, our method significantly performs better on long-tail scenarios. We further conduct ablation studies to highlight the contribution of different proposed components.

## I. INTRODUCTION

Trajectory prediction is a safety-critical task where the goal is to predict the future trajectories of the agents given their history information and the state of their surrounding environment. Relying on such information, the existing trajectory prediction models [1]–[5] achieve promising performance on the benchmarks. However, they are suffering from low accuracy performance on long-tail challenging scenarios.

As a result of the long-tail phenomenon, prediction models focus on more frequent (often simpler) scenarios and tend to put less emphasis on rarer challenging cases [6]. This limits the applicability of the existing approaches to practical intelligent driving systems.

A commonly adopted approach to the long-tail problem in trajectory prediction is employing contrastive learning which aims to better organize latent features for more balanced training [6], [7]. However, this scheme is not compatible with many existing architectures [8]–[11], as they employ multiple encoded vectors in their latent space bottleneck. For example, one for each agent or road element. Moreover, contrastive learning can impose additional computational burden which is not desirable for practical systems.

Alternatively, the long-tail problem can be addressed using multiple specialized experts, each focusing on a particular sub-task [12]. This allows the model to equally pay attention to each subset of the data regardless of their distribution. A shortcoming of this solution, however, is the way the experts are aggregated which adds to computational overhead.

To this end, we propose a novel framework **AMEND: A Mixture of Experts Framework for Long-tailed Trajectory Prediction**. Our framework is based on the divide-and-conquer technique, where a complex task can be decomposed into a set of simpler sub-tasks corresponding to sub-domains in the input space. For example, motion behaviour at intersections is very different from that on straight roadways.

Our approach follows a two-step training regiment. In the first phase, we cluster the data into distinct sections with shared characteristics and then train an expert model on each cluster. In the next phase, we train a router network by ranking the performance of experts on the training data. During inference time, the router network scores each expert given the test sample, and based on the score, a selection module chooses which expert to use to generate predictions.

By directing the input to a single expert at a time, our approach avoids any additional computational cost during inference. In addition, our method is model-agnostic and modular, meaning that it treats the backbone model as black-box and only controls its inputs.

Our **contributions** are as follows: We propose a novel mixture of experts framework for trajectory prediction. Our framework encourages the diversity of expert skill sets to mitigate the long-tailed distribution problem, while simultaneously avoids additional computational cost. We conduct empirical evaluations on common pedestrian trajectory benchmark datasets and show that our proposed approach achieves state-of-the-art performance. We conduct additional experiments, highlighting the advantage of our multi-expert method on predicting challenging scenarios. At the end, we perform ablation studies, showing the advantage of each proposed approach on the overall performance.

## II. RELATED WORK

### A. Trajectory Prediction

Trajectory prediction models aim to forecast the future positions of agents given their past trajectories and their surrounding context. There is a large body of literature in this domain, many of which are catered to pedestrian trajectory prediction [1], [10], [11], [13]–[15]. These models rely on variety of architectures, such as recurrent networks [10], [16], [17], graph neural networks [15], [18], [19],

<sup>1</sup> University of Toronto. Work done while at Huawei. ray.mercurius@mail.utoronto.ca

<sup>2</sup> University of Alberta, eahmadi@ualberta.ca

<sup>3</sup> Noah’s Ark Laboratory, Huawei, Canada. first.last@huawei.com

and transformers [1], [11], [13], [14] to effectively capture the complex contextual information. In this work we use Trajectron++ EWTA [6] as our baseline, which is a variation of [15], a graph-based model.

### B. Long-Tailed Learning

Long-tailed learning seeks to improve the performance on tailed samples in imbalanced datasets and it is well studied in the computer vision domain [20]. The re-balancing methods either oversample or undersample imbalanced classes, reweigh the loss function during training, or directly adjust the classification logits during inference to encourage the model to predict low-frequency classes [21]–[23]. A shortcoming of re-balancing methods is that they only perform sample removal or duplication, without adding any new information. This issue is resolved in information augmentation methods, which create new training examples in the tail classes [24]. These methods, however, work best with low-dimensional and simple data distributions, which are not the case for autonomous driving data [24].

The long-tail problem has also been investigated in trajectory prediction. The authors of [6] utilise contrastive learning to separate the difficult scenarios from the easy ones in the latent space allowing the model to better recognize and share information between difficult scenarios. FEND [7] improves the contrastive learning framework by introducing artificial classes formed by clustering the encoded feature vectors of an autoencoder network. A shortcoming of these techniques is that they work with a single latent vector that captures all the scene information. State-of-the-art trajectory prediction architectures [8], [9] employ multiple latent vectors at the bottleneck, such as one for each scene object, and therefore there is no singular feature vector to reshape according to the sample’s class.

FEND used a class-conditioned hypernetwork [25] decoder that allows dynamic and specialized decoder weights for different scenario types [7]. However hypernetworks have many limitations, such as challenges in parameter initialization and complex architectures that must follow. In this work we propose a model-agnostic framework that relies on multiple experts in a computationally efficient fashion without the need for contrastive learning.

### C. Mixture of Experts

Mixture of Experts (MoE) is a machine learning technique that utilizes several base learners, each one specialized on a particular sub-task [26]. MoE differs from ensembling in that only one or a few experts are run for each input value and it can be restricted to only a portion of the model’s architecture [27]. MoE is very effective in increasing accuracy without proportional increase in the computational cost.

MoE has been applied to various sequence analysis tasks, such as natural language processing [26], [28]. Of interest, the approach proposed in [29] uses a novel routing algorithms, where instead of each input being routed to the top- $k$  experts, each expert selects its top- $k$  inputs. In our work, we adopt a routing network that is trained to score the experts

based on the input sample, and in turn uses the best expert to generate trajectory output.

## III. PROBLEM FORMULATION

Given input information consisting of trajectory histories of  $N$  agents in the scene  $x_t^i \in \mathbb{R}^2$ ,  $i \in [1, N]$ ,  $t \in [1 - T_{hist}, 0]$ , where  $x_t^i$  is the 2D coordinates of agent  $i$  at timestep  $t$  and  $T_{hist}$  is the number of history timesteps, our task is to predict the agents’ future trajectories  $y_t \in \mathbb{R}^2$ ,  $t \in [1, T_{pred}]$ , where  $y_t$  is the coordinates of the agents at time  $t$  and  $T_{pred}$  is the number of prediction timesteps.

## IV. METHODOLOGY

In this section, we describe our proposed solution to address the long-tailed learning problem in trajectory prediction. An overview of the proposed method is shown in Figure 1. Individual parts of the method are described below.

### A. Training Specialized Experts

Our objective is to train multiple experts on a diverse dataset exhibiting a long-tailed distribution, assigning each expert to concentrate on distinct data patterns. By segmenting the learning task into simpler uniform sub-tasks, we facilitate more effective learning for each expert. Due to the lack of explicit labels, we will employ unsupervised learning strategies to segregate the dataset into sub-tasks.

We divide the original dataset  $D$  into  $C$  mutually exclusive subsets called  $D_{1:C}$ . We assign a unique expert to each subset to create  $C$  experts represented by  $E_{1:C}$ . The samples within each subset should be adequately similar to allow expert specialisation.

**Clustering in latent space.** Processing and clustering high dimensional inputs is challenging, hence, we use an encoder network and cluster the dataset based on the feature encodings. We select our encoder to be that of Trajectron++ EWTA [6] trained on the same dataset. We chose the encoder of a trajectory prediction model as it naturally embeds input information for a purpose that aligns with our final end-goal of forecasting trajectories.

Effective encoders only keep information relevant for the training task and map similar samples nearer in the latent space. Therefore, our latent space clusters should contain scenarios similar from a trajectory prediction standpoint. To achieve this, we perform K-means clustering on the latent vector of the encoder. Our approach differs from previous sample clustering methods for trajectory prediction, such as FEND [7], which forms clusters on the latent space of an autoencoder applied to individual trajectories. We include additional contextual information in our latent space, such as nearby agent behaviour.

**Loss function.** Our goal is to force the model during training to focus more on specific subdomains of data containing unique prediction patterns, but without losing generalization. To satisfy this trade-off, we train experts on all samples using a modified loss function that assigns more weight to samples

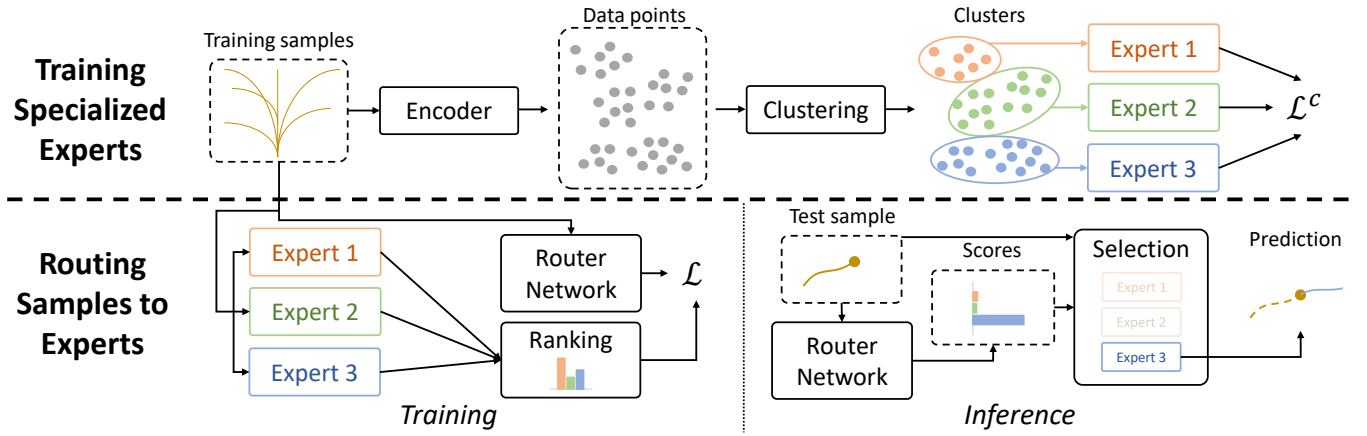


Fig. 1: Overview of the proposed approach. We cluster the data samples based on the latent vector of an encoder network. During training of the experts the loss function is adjusted so that each expert focuses on a particular sample cluster. Next, we calculate the relative performance rankings of the experts, which are used to generate targets to train the router network. At inference we use the router network to select best expert to generate the predictions.

belonging to the expert’s assigned cluster. The training loss function of an expert over a batch is defined as:

$$\mathcal{L}^c = \frac{1}{B} \sum_{i=1}^B (\mathbb{1}[x_i \in D_c](1 + \alpha) + \mathbb{1}[x_i \notin D_c](1 - \alpha)) \mathcal{L}_i^c, \quad (1)$$

where  $B$  is the batch size,  $x_i$  is a training example,  $c$  denotes a particular expert,  $\mathcal{L}_i^c$  is the original loss of expert  $E_c$  on sample  $x_i$  in our baseline model,  $\mathbb{1}(R)$  is the identity function that returns 1 when the condition  $R$  is satisfied and otherwise 0, and  $\alpha$  is a hyperparameter. Setting  $\alpha = 1$  results in mutually exclusive input data subsets for the experts, while setting  $\alpha = 0$  results in identical training.

### B. Routing Samples to Experts

During inference, our challenge is how to assign the test samples to the best expert to perform the forward pass and generate predictions. We propose to use a router network to predict the aptitude of each expert on a given test sample. The router network takes in the input information and outputs a confidence score  $p_{1:C}$  for each of the  $C$  experts, where  $\sum_{c=1}^C p_c = 1$ , indicating the probability of that expert being the best expert for the given sample. Then, the expert with the highest confidence score is selected. The architecture of the router network consists of an encoder network (adopted from the baseline) followed by two fully connected layers.

For router training, to identify the best-performing expert for a given sample, we rely on Average-Displacement-Error (ADE) and Final-Displacement-Error (FDE) metrics,

$$c_{best} = \underset{c}{\operatorname{argmin}} R_{FDE}^c + R_{ADE}^c, \quad (2)$$

where  $c_{best}$  indicates the best expert that has the lowest combined FDE and ADE among all experts, and  $R_{FDE}^c, R_{ADE}^c \in \mathbb{N}$  are the rankings of the expert  $E_c$  on the  $FDE$  and  $ADE$  metrics, respectively, with  $R = 1$  indicating the best performing expert. We use cross-entropy loss to train the router network, with the target being a one-hot vector where index  $c_{best}$  is one.

At inference step, we direct the inputs to the expert with the highest confidence score in a winner-takes-all aggregation scheme. An advantage of this is that the forward pass is only computed for a single expert, hence, the inference computational cost does not scale with the number of experts.

## V. EXPERIMENTS

### A. Experimental Setup

**Datasets.** We evaluate the models on ETH-UCY, which are bird’s-eye-view pedestrian benchmark datasets [36], [37]. The datasets contain challenging scenarios, such as crowded scenes with complex agent-to-agent interactions. Following the previous works [1], [11], [15], we average our final results over a five-fold cross-validation scheme, with four splits utilized for training and one for test. For the ETH-UCY datasets,  $T_{pred}$  is 12,  $T_{hist}$  is 8 and the samples are collected at  $2.5Hz$  ( $\Delta t = 0.4s$ ).

**Metrics.** We use the common performance metrics for trajectory prediction [1], [11], [15], namely Average-Displacement-Error (ADE) and Final-Displacement-Error (FDE), and report the minimum error across  $K = 20$  predictions.

For evaluating performance on tail samples, we use the following methods:

i) *Scenario Difficulty Ranking:* We evaluate the model’s performance on the top 1% and 5% difficult scenarios. To rank the scenarios by difficulty we utilise the errors of a simple Kalman filter [6].

A drawback of this metric is that the definition of the tailed scenarios is dependant on the model used to judge difficulty. Therefore, different models might underperform on different scenarios and the definition of the data tail might vary [7]. Furthermore, simply measuring errors on the set of challenging scenarios does not properly capture the changes in the distribution of errors across the dataset. It is possible for a trade-off to occur in which the model’s performance deteriorates on other scenarios. ii) *Error Distribution Quantiles:* We use an alternative metric which directly measures the magnitude of the tail of the error distribution. This is

TABLE I: Quantitative evaluation on the UCY/ETH benchmark. Results are reported on minADE<sub>20</sub>/minFDE<sub>20</sub>. **Bold** numbers indicate the best performance for each metric, underline indicates the second best performance for each metric. For both metrics, lower value is better.

Method	ETH	Hotel	Univ	Zara1	Zara2	Avg
Traj++ EWTA [6]	0.38/0.65	0.10/0.17	0.18/0.34	0.14/0.27	0.10/0.19	0.18/0.32
MID [30]	0.39/0.66	0.13/0.22	0.22/0.45	0.17/0.30	0.13/0.27	0.21/0.38
Leapfrog [31]	0.39/0.58	0.11/0.17	0.26/0.43	0.18/0.26	0.13/0.22	0.21/0.33
NSP-SFM [32]	<b>0.25/0.24</b>	<u>0.09/0.13</u>	0.21/0.38	0.16/0.27	0.12/0.20	0.17/0.24
Y-Net [33]	0.28/ <u>0.33</u>	0.10/0.14	0.24/0.41	0.17/0.27	0.13/0.22	0.18/0.27
SIC-MTP [34]	<u>0.27/0.45</u>	0.11/0.16	0.26/0.46	0.19/0.33	0.13/0.26	0.19/0.33
EQMotion [35]	0.40/0.61	0.12/0.18	0.23/0.43	0.18/0.32	0.13/0.23	0.21/0.35
AMEND (Ours)	0.29/0.42	<b>0.08/0.12</b>	<b>0.13/0.21</b>	<b>0.12/0.20</b>	<b>0.09/0.14</b>	<b>0.14/0.22</b>

TABLE II: Quantitative evaluation on long-tail scenarios for the ETH-UCY benchmark computed based on the weighted average of a five-fold evaluation. Results are reported on minADE<sub>20</sub>/minFDE<sub>20</sub>. Top  $\alpha\%$  split refers to performance on the highest percentile of challenging scenarios. VaR<sub>99</sub> and VaR<sub>95</sub> refer to the 0.99 and 0.95 quantiles of the error distribution. Relative metrics are provided in the last two columns as a normalized measure of the model performance on the long-tail. **Bold** numbers indicate the best performance for each metric. For all metrics lower value is better.

Method	minADE <sub>20</sub> /minFDE <sub>20</sub>					$\frac{\text{minADE}_{20}}{\text{minADE}_{20}^{All}} / \frac{\text{minFDE}_{20}}{\text{minFDE}_{20}^{All}}$	
	Top 1 %	Top 5 %	All	VaR <sub>99</sub>	VaR <sub>95</sub>	Top 1%	Top 5%
Traj++ [15]	0.58/1.23	0.65/1.42	0.21/0.41	0.98/2.47	0.56/1.33	2.8/3.0	3.1/3.5
Traj++ EWTA [6]	0.45/0.88	0.44/0.91	0.18/0.32	0.75/1.92	0.47/0.92	2.5/2.8	2.4/2.8
contrastive [6]	0.40/0.71	0.44/0.91	0.19/0.32	0.74/1.82	0.47/0.93	2.1/2.2	2.3/2.8
FEND [7]	0.38/0.74	0.37/0.76	0.17/0.32	-	-	2.2/2.3	2.2/2.4
AMEND(Ours)	<b>0.27/0.44</b>	<b>0.25/0.41</b>	<b>0.14/0.22</b>	<b>0.63/1.53</b>	<b>0.36/0.67</b>	<b>1.9/2.0</b>	<b>1.8/1.9</b>

the error on the worst performing samples according to the model. We adopt the value-at-risk (VaR) metric. VaR <sub>$\alpha$</sub>  refers to the  $\alpha^{th}$  quantile of the error distribution, where  $\alpha \in (0, 1)$ :

$$\text{VaR}_{\alpha}(E) = \inf\{e \in E : P(E \geq e) \leq 1 - \alpha\}. \quad (3)$$

More formally, it is the smallest error  $e$  such that the probability of observing error larger than  $e$  is smaller than  $1 - \alpha$ , where  $E$  is the distribution of errors. We measure VaR at three levels: 0.95, 0.97 and 0.99.

**Baseline.** We used Trajeon++ EWTA (Traj++ EWTA for short) [6] as the baseline model for our framework. Traj++ EWTA enhances Traj++ [15] by replacing its conditional Variational Auto Encoder (CVAE) module with the multi-hypothesis networks trained with Evolving Winner-Takes-All (EWTA).

**Implementation Details.** Our main model and the router network follow the same training schedule. We train our model for 100 epochs consisting of 20 epochs for each value of the  $K \in \{20, 10, 5, 2, 1\}$  in EWTA stages. The router softmax temperature is set to 1. The dimension of our router decoder’s hidden layer is 232. All other training details are kept the same as in the baseline model [6].

**Data Preprocessing.** To eliminate the impact of arbitrary scale, we normalize the dimensions similar to [6]. The trajectory coordinates are divided by their standard deviation in the training dataset. Additionally, we normalize the headings by rotating the inputs so that the last known direction of the agent points in the positive y-axis. The opposite rotation is then applied to the model’s outputs. With this orientation normalization the trajectory end-points capture the general intention of the agents.

## B. Experimental Variations

**Clustering on trajectory endpoints.** We experiment with modifying the clustering algorithm used to partition the dataset. Instead of clustering on the latent space, we perform K-Means [38] clustering on endpoints of the ego-vehicle’s future trajectories. We denote the model that uses this clustering as Trajectory. Empirically, we find that this results in partitions based on modality, such as turn type and velocity. A shortcoming is that it only utilises information from the ego’s trajectory, and ignores other scene information such as interactions with nearby agents or potential map info.

**Cluster Assignment.** In this experiment we replace the router network with a heuristic algorithm to generate expert confidence scores. We rely on the principle that an expert should perform best on examples most similar to its assigned training cluster. We utilize the distance between the embedding of two arbitrary samples in the latent space as a proxy for how similar the samples are. Therefore, given a sample, we calculate the latent distance between itself and the cluster centroid of each expert as a rough approximation to the confidence of each expert on that sample. Recall that each expert  $E_{1:C}$  focused on a unique subset of data  $D_{1:C}$  during training, and is associated with a unique cluster centroid denoted by  $\phi_{1:C}$ . Confidence scores are generated via a Softmax operation on these distances as follows:

$$p_i^c = \frac{\exp(-\text{dist}(\rho(x_i), \phi_c))}{\sum_{j=1}^C \exp(-\text{dist}(\rho(x_i), \phi_j))} \quad (4)$$

where  $i$  denotes sample index,  $c$  denotes the expert index,  $\rho(\cdot)$  is our encoder network and  $\phi_j$  is a cluster centroid. We

denote this model as Cluster-based.

### C. Comparison to SOTA Predictors

In Table I, we compare, our proposed approach (AMEND) to several state-of-the-art pedestrian trajectory prediction models on the ETH-UCY datasets. Though our framework is designed to improve performance on the tail of the dataset, we observe that it significantly improves performance overall. This is because we effectively scaled our baseline model, increasing the number of expert models and giving each specialized abilities. Compared to the baseline, we improved performance by 22% and 31% on minADE and minFDE, respectively. Compared to SOTA models, we achieve the best performance on most subsets, and the best performance overall by up to 18%.

### D. Long-tailed Prediction

We compare our method to trajectory prediction models catered to long-tailed prediction. All models utilize variations of Trajectron++ [15] as their backbone model which provides a fair comparison. As shown in Table II, our framework surpasses others on long-tail evaluation metrics by a significant margin. When compared to the runner-up FEND [7], we achieve improvements of 29% and 41% in minADE and minFDE on top 1% consisting of the most difficult scenarios. Across all scenarios, we achieve 18% and 31% improvement on minADE and minFDE errors, respectively.

From the relative metrics (the last two columns), we can see that given our higher performance ratio on challenging scenarios, compared to the average case, our model achieves a more balanced performance across different difficulty levels. This indicates that our approach, which consists of allocating more training resources to specific prediction patterns via specialized modules is especially helpful for complex patterns. We verify the soundness of our improvements by comparing the VaR metric which measures the error distribution quantiles. Achieving lower values on VaR metrics overall means that the largest prediction errors given by our model across the dataset is smaller than the largest errors of other models. This indicates that we have not migrated errors from one domain to another one.

### E. Training Diverse Experts

In Figure 2, we show the performance of each experts on test samples across different clusters. As expected, the experts perform best in the cluster of scenarios assigned to them during training. Note that the best performing expert for each cluster tends to be the expert that was assigned to that cluster.

In Table III, we compare different training methods to create specialized experts. For clustering (on top) we compare Trajectory, which is a model that clusters scenarios based on the final endpoints of the ego-trajectory to ours (Latent Feature). Here, we can see that our approach marginally outperforms Trajectory, especially on top 5% samples. Such improvement can be due to added information captured in the latent space, accounting for factors, such as interactions between the agents.

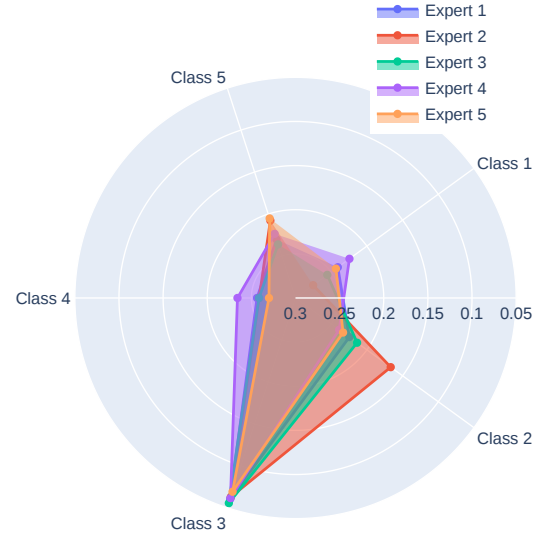


Fig. 2: Radar plot showing the performance of each expert on the different cluster splits of the ETH-UCY Hotel test dataset showing significant variations in the performance of the experts. The best performing expert for each cluster tends to be the one that was assigned to it during training.

TABLE III: The effect of clustering basis (top rows) and the routing method (bottom rows) on the overall performance and the long-tailed performance of the AMEND model. The minADE<sub>20</sub>/minFDE<sub>20</sub> are reported based on the average of 5-fold evaluation for the ETH/UCY dataset. The \* indicates the default setting.

	Top 1 %	Top 5 %	All	VaR <sub>97</sub>
Trajectory	0.28/0.42	0.27/0.45	0.14/0.22	0.44/0.91
Latent Feature*	0.27/0.44	0.25/0.41	0.14/0.22	0.44/0.91
Cluster-based	0.26/0.42	0.26/0.44	0.14/0.22	0.45/1.01
Router Network*	0.27/0.44	0.25/0.41	0.14/0.22	0.44/0.91

### F. Effectiveness of Router Network

In Table III (bottom) we compare our main model which selects experts with the router network (Router Network), to an alternative approach which selects the expert who's sample cluster is predicted to contain the test sample (Cluster-based). Here, we can see that utilising the router network generates better results, especially on the VaR metric where improvement of up to 10% is achieved.

We further investigate the discrepancy between the performance of our routing approach compared to the clustering technique. For this experiment, we report the results in terms of accuracy of the methods for predicting which expert would perform best. The results are summarized in Table IV. Here, the baseline for comparison is random routing, which has a 20% chance of being correct since we have 5 experts. The strong performance of Cluster-based relative to random routing supports the hypothesis that the best performing expert for samples within a particular cluster is usually the expert specialized for the cluster. However, its lower performance compared to our Router-Network approach, suggests that there are exceptions to this rule, which our router has successfully learned. This supports the idea of using neural networks to map the complex distribution of relative expert strength across the data space for routing.

TABLE IV: The accuracy of routing method in selecting the expert that performs best on a test sample. Best performing expert is defined as the one that achieves the lowest error metrics (ADE/FDE). Errors per sample are averaged over sixteen trials to reduce uncertainty. The \* indicates our default approach. Higher value means better.

	Random	Cluster-based	Router Network*
ETH	-	0.32/0.37	0.35/0.36
Hotel	-	0.39/0.31	0.38/0.38
Univ	-	0.32/0.31	0.38/0.37
Zara1	-	0.28/0.29	0.31/0.32
Zara2	-	0.48/0.38	0.49/0.41
Avg	0.20/0.20	0.36/0.33	0.38/0.37

## VI. CONCLUSION

In this paper, we tackled the long-tail pedestrian prediction problem by formulating a Mixture of Experts framework. We proposed a novel two-stage training scheme in which we first train specialized experts on sub-tasks within the data, and second use the experts to train a routing network for scoring the experts at inference time. We demonstrated that clustering the dataset and focusing each expert’s resources on a partition creates specialized skills which can be utilised to generate accurate predictions. We conducted extensive experimental evaluation on common pedestrian trajectory benchmark datasets and showed that our approach achieves state-of-the-art performance while outperforming the previous methods on challenging tailed samples. We further highlighted the effectiveness of our proposed modules via ablation studies.

## REFERENCES

- [1] Y. Yuan, X. Weng, Y. Ou, and K. M. Kitani, “Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting,” in *ICCV*, 2021.
- [2] M. Pourkeshavarz, C. Chen, and A. Rasouli, “Learn tarot with mentor: A meta-learned self-supervised approach for trajectory prediction,” in *ICCV*, 2023.
- [3] D. Zhu, G. Zhai, Y. Di, F. Manhardt, H. Berkemeyer, T. Tran, N. Navab, F. Tombari, and B. Busam, “Ipcc-tp: Utilizing incremental pearson correlation coefficient for joint multi-agent trajectory prediction,” in *CVPR*, 2023.
- [5] E. Amirloo, A. Rasouli, P. Lakner, M. Rohani, and J. Luo, “Latentformer: Multi-agent transformer-based interaction modeling and trajectory prediction,” *arXiv preprint arXiv:2203.01880*, 2022.
- [6] O. Makansi, O. Cicek, Y. Marrakchi, and T. Brox, “On exposing the challenging long tail in future prediction of traffic actors,” in *ICCV*, 2021.
- [7] Y. Wang, P. Zhang, L. Bai, and J. Xue, “FEND: A future enhanced distribution-aware contrastive learning framework for long-tail trajectory prediction,” in *CVPR*, 2023.
- [8] S. Shi, L. Jiang, D. Dai, and B. Schiele, “Motion transformer with global intention localization and local movement refinement,” in *NeurIPS*, 2022.
- [9] Y. Gan, H. Xiao, Y. Zhao, E. Zhang, Z. Huang, X. Ye, and L. Ge, “MGTR: Multi-granular transformer for motion prediction with lidar,” *arXiv preprint arXiv:2312.02409*, 2023.
- [10] A. Rasouli, M. Rohani, and J. Luo, “Bifold and semantic reasoning for pedestrian behavior prediction,” in *ICCV*, 2021.
- [11] L. Shi, L. Wang, S. Zhou, and G. Hua, “Trajectory unified transformer for pedestrian trajectory prediction,” in *ICCV*, 2023.
- [12] Y. Zhang, B. Hooi, L. Hong, and J. Feng, “Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition,” in *CVPR*, 2022.
- [13] A. Rasouli and I. Kotseruba, “Pedformer: Pedestrian behavior prediction via cross-modal attention modulation and gated multitask learning,” in *ICRA*, 2023.
- [4] R. Karim, S. M. A. Shabestary, and A. Rasouli, “Destine: Dynamic goal queries with temporal transductive alignment for trajectory prediction,” *arXiv preprint arXiv:2310.07438*, 2023.
- [14] A. Rasouli, “A novel benchmarking paradigm and a scale-and motion-aware model for egocentric pedestrian trajectory prediction,” *arXiv preprint arXiv:2310.10424*, 2023.
- [15] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, “Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data,” in *ECCV*, 2020.
- [16] Z. Su, S. Zhang, and W. Hua, “CR-LSTM: Collision-prior guided social refinement for pedestrian trajectory prediction,” in *IROS*, 2021.
- [17] P. Dendorfer, S. Elflein, and L. Leal-Taixe, “MG-GAN: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction,” in *ICCV*, 2021.
- [18] A. Hasan, P. Sriram, and K. Driggs-Campbell, “Meta-path analysis on spatio-temporal graphs for pedestrian trajectory prediction,” in *ICRA*, 2022.
- [19] L. Shi, L. Wang, C. Long, S. Zhou, M. Zhou, Z. Niu, and G. Hua, “SGCN: Sparse graph convolution network for pedestrian trajectory prediction,” in *CVPR*, 2021.
- [20] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, “Deep Long-Tailed Learning a survey,” in *PAMI*, 2023.
- [21] A. Estabrooks, T. Jo, and N. Japkowicz, “A multiple resampling method for learning from imbalanced data sets,” in *Computational Intelligence*, 2004.
- [22] Z.-H. Zhou and X.-Y. Liu, “Training cost-sensitive neural networks with methods addressing the class imbalance problem,” in *TKDE*, 2005.
- [23] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, “Long-tail learning via logit adjustment,” in *ICLR*, 2021.
- [24] D. Rempe, J. Philion, L. J. Guibas, S. Fidler, and O. Litany, “Generating useful accident-prone driving scenarios via a learned traffic prior,” in *CVPR*, 2022.
- [25] D. Ha, A. M. Dai, and Q. V. Le, “Hypernetworks,” in *ICLR*, 2017.
- [26] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, “Adaptive mixtures of local experts,” *Neural Computation*, vol. 3, no. 1, pp. 79–87, 1991.
- [27] S. E. Yuksel, J. N. Wilson, and P. D. Gader, “Twenty years of mixture of experts,” in *TNNLS*, 2012.
- [28] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer,” in *ICLR*, 2017.
- [29] Y. Zhou, T. Lei, H. Liu, N. Du, Y. Huang, V. Y. Zhao, A. M. Dai, Z. Chen, Q. V. Le, and J. Laudon, “Mixture-of-experts with expert choice routing,” in *NeurIPS*, 2022.
- [30] T. Gu, G. Chen, J. Li, C. Lin, Y. Rao, J. Zhou, and J. Lu, “Stochastic trajectory prediction via motion indeterminacy diffusion,” in *CVPR*, 2022.
- [31] W. Mao, C. Xu, Q. Zhu, S. Chen, and Y. Wang, “Leapfrog diffusion model for stochastic trajectory prediction,” in *CVPR*, 2023.
- [32] J. Yue, D. Manocha, and H. Wang, “Human trajectory prediction via neural social physics,” in *ECCV*, 2022.
- [33] K. Mangalam, Y. An, H. Girase, and J. Malik, “From goals and waypoints & paths to long term human trajectory forecasting,” in *ICCV*, 2021.
- [34] Y. Dong, L. Wang, S. Zhou, and G. Hua, “Sparse instance conditioned multimodal trajectory prediction,” in *ICCV*, 2023.
- [35] C. Xu, R. T. Tan, Y. Tan, S. Chen, Y. G. Wang, X. Wang, and Y. Wang, “EqMotion: Equivariant multi-agent motion prediction with invariant interaction reasoning,” in *CVPR*, 2023.
- [36] S. Pellegrini, A. Ess, K. Schindler, and L. van Gool, “You’ll Never Walk Alone modeling social behavior for multi-target tracking,” 2009.
- [37] L. Alon, C. Yiorgos, Lischinski, and Dani, “Crowds by example,” *Computer Graphics Forum*, vol. 26, no. 3, pp. 655–664, 2007.
- [38] J. MacQueen *et al.*, “Some methods for classification and analysis of multivariate observations,” in *The Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, no. 14, 1967, pp. 281–297.