scientific reports

Check for updates

OPEN

Trajectory prediction via proposal guided transformer with out way attention

Wei Xu, Ruochen Li, Xiaodong Du, Bingjie Li & Lei Xing✉

The accurate prediction of the behavior of surrounding agents is crucial for the safe operation of autonomous vehicles. Currently, the dominant approach involves manually defining rules, which often fail to cover all potential scenarios. To address the rigidity and challenges of generalizing these rules to real-world driving contexts, we introduce a novel attention mechanism called "Out Way Attention". This mechanism improves the model's capacity to dynamically adapt to various driving situations by incorporating attention exits (out way) into the attention framework. Additionally, we present a new trajectory prediction framework that includes a learnable proposal matrix and permutation-invariant positional encoding. This matrix aids in forecasting future multimodal trajectories for multiple interacting agents in dynamic settings. The permutation-invariant positional encoding ensures that the processing sequence of agents at the same time does not influence the prediction outcomes. By integrating the proposed methods into the Transformer architecture, our approach reduces human intervention and significantly enhances the model's adaptability, as well as improving training and inference efficiency. We have validated the effectiveness of our model on three public datasets, Argoverse, Trajnet++, and ETH/UCY. The results confirm that our method sustains robust performance in complex, highly dynamic environments with multiple interacting agents.

Keywords Out Way Attention, Trajectory prediction, Learnable proposal matrix, Transformer

The rapid advancement in computer and communication technologies has greatly enhanced autonomous driving, particularly through the development of trajectory prediction systems. These systems are crucial for autonomous vehicles to anticipate future road conditions and make informed decisions, thereby improving safety and efficiency[1]. Trajectory prediction involves forecasting the future positions of vehicles based on the analysis of motion patterns, a task complicated by the dynamic and unpredictable nature of driving environments[2,3]. Early trajectory prediction models treated the task as deterministic, providing a single possible future path. However, this approach often fails to capture the range of potential outcomes in highly variable traffic conditions[4]. Recent advancements have shifted towards multi-modal trajectory prediction, which considers multiple possible future trajectories. This shift helps address the limitations of the unimodal models by preventing mode collapse and enhancing the robustness and interpretability of the predictions[5–10]. In the current prediction task, many researchers have made efforts in capturing the interaction between different agents[11] and inducing multimodal trajectory generation mode problems[12–14].

However, many current studies focus solely on improving prediction accuracy, concentrating on specific traffic scenarios. For example, when considering the neighboring agents of the ego-agent, they either set thresholds[5] such as the range or number of agents or directly set weights for longer distances[15]. The problem is that these artificially customized rules often overlook certain scenarios. To address this issue, we propose an improved Out Way Attention mechanism to reduce boundary condition constraints and gradually generalize to more closely resemble real-world environments. By incorporating attention exits (out way) into the attention mechanism, the model can gradually reduce the focus on agents that do not require attention, such as vehicles at greater distances. This allows the entire model to be trained in an end-to-end manner.

Previous multimodal trajectory prediction methods often required multiple predictions to forecast various trajectories. However, our introduction of a learnable proposal matrix enables the model to output all possible predicted trajectories in a single forward pass, significantly enhancing both training and inference efficiency.

The original attention mechanism requires the use of position encoding to introduce temporal information into the attention process[16]. However, in the complex task of trajectory prediction, where multiple agents exist in different states at the same moment, the processing order of agents at the same moment should not affect

College of Transportation, Shandong University of Science and Technology, Qingdao 266590, China. ✉email: xinglei0915@163.com

the prediction outcomes. Our improved permutation-invariant positional encoding ensures the permutation invariance of agents.

Based on the improved attention mechanism, we propose a multi-modal trajectory training strategy called PowFormer. Using the latent variable sequences, it effectively utilizes the scene and interaction environment to output multiple future trajectories of traffic participants.

Overall, compared to current state-of-the-art methods[17], the main advantage of our approach is that it can output multiple trajectories simultaneously by using latent variables, which significantly improves prediction efficiency. Compared to the method in[18], our model is based on Transformer and out-way attention, offering higher model capability and achieving better performance. Compared to PPT[19], which enhances the model's ability to capture short-term dynamics and long-term dependencies through gradual training, our method is more straightforward. It does not require introducing prior knowledge of long or short term dependencies but instead leverages attention mechanisms to autonomously learn and enhance the model's capability.

Finally, we validate our model using the autonomous driving dataset Argoverse[20] and the pedestrian dataset Trajnet++[21] and ETH/UCY. The evaluation in different scenarios demonstrates that our method achieves the best results across multiple datasets and exhibits excellent generalization ability.

In summary, the main contributions of this paper are as follows:

- We introduce an enhanced Out Way Attention mechanism that reduces the reliance on boundary conditions, fading the influence of distant vehicle interactions, and avoids the need for manually crafting cumbersome rules.
- Our proposal of a learnable proposal matrix facilitates the simultaneous output of all potential trajectories in a single forward pass, drastically improving the model's efficiency during training and inference phases.
- We implement an improved permutation-invariant positional encoding within the attention mechanism, ensuring that the simultaneous processing of multiple agents does not bias the trajectory prediction outcomes.

Related works

Motion prediction

Existing approaches typically treated motion prediction as a deterministic problem. From the perspective of output trajectory quantity, they usually only output the most probable trajectory or use the mean as the prediction result. This method, known as determin- istic trajectory prediction(DTP)[6], cannot satisfy the demand for high-precision predictions. To address the uncertainty in trajectory prediction, fully reflect the dynamics of vehicle operation, and maintain agility in the face of potential dangers, many advanced approaches have been proposed multi-modal trajectory prediction (MTP) methods. This work is challenging because only one possibility exists in each training sample.

The simplest way to convert DTP to MTP is through Gaussian trajectory prediction based on noise. Gupta[12] first introduced the GAN-based MTP framework, which combines encoded historical information with random noise and then inputs it into the decoding module. Some researchers have introduced this method into architectures with explicit encoding methods[22,23], and methods based on normalized flows perform reversible transformations of parameters through reversible networks. However, training strategies based on noise are limited by a predefined two-dimensional shape, and the strong randomness and continuity are insufficient to simulate discrete possibilities. Recently, many studies have achieved multi-modal trajectory prediction through prediction methods based on anchor points. Some methods use an endpoint-based approach to model the intermediate path based on the historical trajectory and endpoint. TNT[13] heuristically sets multiple possible locations, MultiPath[14] obtains potential endpoints from anchor point trajectories through clustering, and CoverNet[24] views the trajectory prediction problem as a classification problem under discrete trajectory clusters. Compared with methods based on noise, this approach does not limit uncertainty to a fixed form and reflects the characteristics of uncertainty. DenseTNT[25] directly predicting the probability distribution of endpoints, addressing the problems in TNT[13]. This approach avoide the heuristic setting of anchors, and achieving anchor-free. Wu et al.[26] expanded MultiPath, proposing to generate a cross-distribution of two agents along all trajectory anchor points. However, this explicit encoding method has a complex process, and when the positions of neighbour agents change significantly, the ego-agent cannot quickly generate new trajectories based on established anchors. Although the explicit encoding mode of multi-modal trajectory generation induced by anchor points has strong interpretability, it limits the trajectory to predefined points, has a strong dependence on anchor point quality, and needs to improve its flexibility.

In contrast, our method of capturing hidden intentions with learnable proposal matrix to achieve multi-modal trajectories has strong flexibility. Moreover, our method models the future trajectories of all agents and jointly predicts trajectories with a sense of social interaction.

Social interaction modeling

Social interactions have a significant impact on the future movement of the agent. Social interaction modeling usually considers the time dimension and the social dimension. The interaction between agents is mostly captured through social pooling[12,27,28], graph neural networks[29], or attention mechanisms[30–33]. Grid-based BEV images are a common method for representing map structures and neighborhood relationships between agents. In the image, the prediction task can be simplified into a trajectory selection and offset problem[34]. By using lane grids[25,35], a probability heatmap is generated in a sparse manner. However, this lane-generated heatmap confines the prediction range to specific drivable areas and fixed lane widths. Through our improved positional encoding and Out Way Attention mechanism, we can simultaneously handle temporal and social features, and still model these two features in different dimensions. However, we break the artificial rules and generalize the scenarios to be closer to real-world situations.

Multimodal trajectory prediction based on transformer

The Transformer[16] is known for its ability to recognize and process long-term complex data, and many models based on the Transformer architecture have emerged in the field of trajectory prediction to capture dependencies in the time and space domains. Huang et al.[7] studied the dynamic interaction between vehicles using a hierarchical game approach and proposed the Gameformer model; mmTransformer[8] designed a network structure based on the stacked transformer to divide the future drivable area; Zhu et al.[9] captured the interaction between vehicles through the future driving intent and behavior of the vehicle and proposed the BiFF model; HiVT[10] decomposed the problem into local context extraction and global interaction modeling and proposed a hierarchical vector Transformer; Zhang et al. proposed MTPT[36] focusing on enhancing the interpretability of the model and improving the attention mechanism; Scene Transformer[37] proposed a unified framework for multi-target trajectories, ensuring the scene consistency of the output trajectories; in Agentformer[15], the temporal and social dimensions are flattened in the form of CVAE, and a sampling mechanism is added to increase the diversity of sampled trajectories, which are described in probabilistic form. In this paper, we use the method of learnable proposal matrix. This method helps the model to capture latent features and improves the model speed compared to the training strategy of decoding through autoregression.

Researchers have proposed a novel Transformer encoders that preserves the critical property of permutation invariance across rows or columns[38]. However, as far as we know, we are the first to incorporate this feature to trajectory prediction.

Approach

An overview of our method is illustrated in Fig. 1. We encode the past trajectories of all agents in the scene through just a single forward computation, and the scene features are shared among all agents to avoid the wastage of computations.

We utilize Multi-Head Out Way Attention to enable the model to focus on the truly important agents, avoiding the need for manually designed rules. In the decoding module, to increase the diversity of sampled trajectories and avoid falling into mode averaging, we introduce a learnable proposal matrix. Each column vector in the matrix corresponds to a driving tendency. We also apply permutation-invariant position encoding in the social dimension and standard position encoding in the temporal dimension to ensure that the order of predicting agents at the same moment does not affect the prediction outcome.We refer to these architectures as "PowFormer". In the following sections, we will provide a detailed introduction to the core modules.

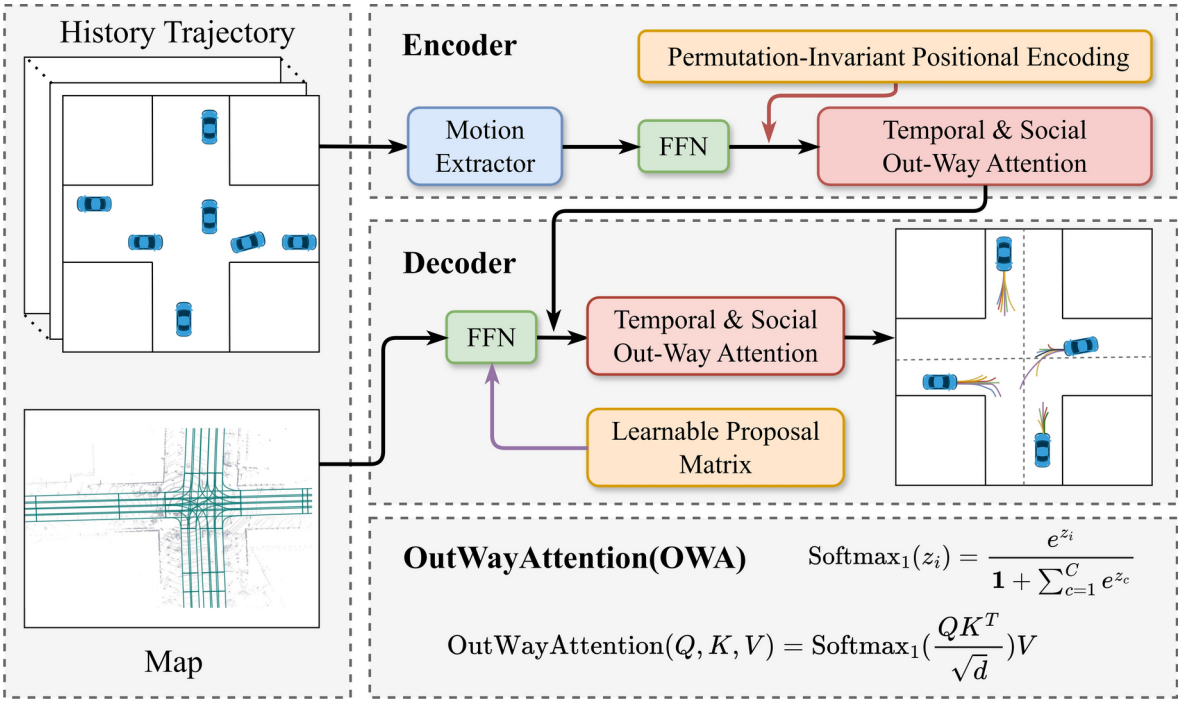

Fig. 1. Overview of the PowFormer: Historical trajectories are extracted through the Motion Extractor, and then integrated with our proposed Permutation-Invariant Positional Encoding, which ensures permutation invariance when computing attention in the spatial dimension, guaranteeing that the output order does not affect the result. Simultaneously, our proposed Out Way Attention mechanism simplifies the previous workflow of specifying thresholds by introducing attention exits (out way), enhancing adaptability in complex environments. Finally, a learnable proposal matrix is introduced in the decoder, allowing us to predict multiple modal trajectories simultaneously, thus improving training and inference efficiency.

Input and output formulation

In this paper, we consider a scene with M agents. The agent whose future trajectory is to be predicted is called the target agent, and the remaining agents are called neighbor agents. There is a set of observation sequences that contain all historical information,

$$X = (X_1, X_2, \cdots, X_{T_{obs}}) \tag{1}$$

$$X_t = \{(x_1, x_2, \cdots, x_K)_1, \cdots, (x_1, x_2, \cdots, x_K)_M\} \tag{2}$$

where $t \in (1, \cdots, T_{obs})$, T_{obs} represents the length of the time sequence observed, and K represents the number of state features contained by the ith agent in this frame. Each agent's state includes time features and spatial features. The spatial features of the ith agent at time t include position, speed, acceleration, heading angle, features of surrounding agents, and the observation information of the past T_{obs} frames. We also input the information captured from the HD map (such as lane polylines, traffic lights, etc.) into the calculation of each frame, which is conducive to reflecting the impact of geographic environment features on the future trajectory of the agent. The predicted trajectory is represented as follows:

$$Y = (Y_{T_{obs}+1}, Y_{T_{obs}+2}, \cdots, Y_{T_{obs+pred}}) \tag{3}$$

where T_{pred} is the length of the predicted time sequence. Given the agent states and map information within an observation window of t time steps, the task of the prediction module is to predict the c future trajectories of M target agents within a range of T_{pred} time steps and calculate the probability of each possible future trajectory.

Out way attention

Out Way Attention builds upon the foundational concept of original Attention[16], articulated as follows:

$$\text{Attention(Q,K,V)} = \text{Softmax}\left(\frac{\text{QK}^\top}{\sqrt{d_k}}\right)\text{V} \tag{4}$$

$$\text{Softmax}\,(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{for } i = 1, 2, \ldots, k \tag{5}$$

$\text{Softmax}\,(z)_i$ is the output probability corresponding to the ith element. z_i is the input score (logit) for the ith class. k is the total number of elements. e is the base of the natural logarithm. This transformation ensures that the outputs are normalized and can be interpreted as probabilities.

The attention score is calculated using three learnable matrices Query, Key, and Value, Q,K,V, d_k is the dimension of the key matrix. In the traffic scene, Q, K and V are projections of the embeddings of past trajectory sequences X in self-attention. In cross-attention inside encoder, Q is the projection of future trajectory sequences Y.

In some cases, the impact of neighboring agents and map features on the driving of the target agent is small. It is not always necessary to pay more attention to closer distances; sometimes, vehicles moving in opposite directions can be ignored. However, when using the original attention mechanism, interactions that can be ignored are also calculated, resulting in a waste of time and possibly failing to focus on the vehicles that truly require attention. This is because the Softmax formula in the attention mechanism has limitations.

The problem of Softmax is that even if there is no related information to add to the output vector, each attention head is still forced to aggregate. We consider a special case where all attention scores in the attention map are extremely small, i.e., the distance between vehicles is very far. They should have extremely small or even close to 0 interaction weights, but Softmax may make the probability of each position become $\frac{1}{k}$.

$$\lim_{x_1 \to -\infty} \cdots \lim_{x_k \to -\infty} \text{Softmax}\,(x)_i = \frac{1}{k} > 0 \tag{6}$$

To solve this problem, Agentformer[15] manually sets a weight of -∞ for vehicles that exceed the threshold. However, the setting of artificial rules increases the complexity of the model, and there may be situations where a vehicle is far away but urgently requires attention. We propose to modify the Softmax formula, simplifying the model's process and making the input more uniform.

$$\text{Softmax}_1(z_i) = \frac{e^{z_i}}{1 + \sum_{c=1}^{C} e^{z_c}} \tag{7}$$

The reason we add a **1** is that we are effectively introducing a virtual object in the attention mechanism, with which any agent computes an attention score of 0, since $e^0 = 1$. However, this virtual object contributes nothing to the output of the attention. The improved Softmax_1 tends to 0 instead of $\frac{1}{k}$ when all x_k tend to -∞, by favoring attention towards this virtual object.

$$\lim_{x_1 \to -\infty} \cdots \lim_{x_k \to -\infty} (\text{Softmax}_1(x))_i = 0 \tag{8}$$

This provides a filtering window for negative infinity and also provides a new option to provide all-low weights, meaning it can choose not to have a high focus on anything.

This new function is used to adjust the original function, where we call the modified attention mechanism Out-Way Attention:

$$\text{OutWayAttention}(\text{Q},\text{K},\text{V}) = \text{Softmax}_1\left(\frac{\text{QK}^\top}{\sqrt{d}}\right)\text{V} \tag{9}$$

Note that we use Out Way Attention to replace the scaled dot-product attention in the original Transformer and still allow multi-head attention to learn distributed representations.

Permutation-invariant positional encoding

The original attention mechanism does not recognize positional information. It is introduced through position encoding.

However, within the same moment, there is no order between agents. If we use basic position encoding[16], the attention mechanism will break the permutation-invariance required. To ensure that the order of agent input at the same moment does not affect the final output, we use a specially designed position encoding. Specifically, different encoding strategies need to be used with the position encoding of time, so we add different position encodings (PE) for the same agent at different moments, and set the same position encoding for all agents at the same moment which is detailed in Fig. 2. The following is the proof.

Definition Let $\text{X} = \{x_1, x_2, \ldots, x_n\}$ be a sequence of embeddings of trajectories (i.e. x is a matrix), each x_i associated with a timestamp t_i. Let $T = \{t_1, t_2, \ldots, t_n\}$ be the corresponding sequence of timestamps. All x_i with the same timestamp t_j are assigned the same positional encoding vector $p(t_j)$.

Setup Consider a special case that $\text{Q} = \text{K} = \text{V} = \text{X} + \text{P}$, where P is the matrix of positional encodings and each row p_i in P corresponds to the positional encoding of x_i, i.e., $p_i = p(t_i)$.

Computation of the attention matrix The attention matrix A is computed as:

$$\text{A} = \text{Softmax}\left(\frac{(\text{X}+\text{P})(\text{X}+\mathbf{P})^\top}{\sqrt{d_k}}\right) \tag{10}$$

Expanding the product yields:

$$(\mathbf{X}+\mathbf{P})(\mathbf{X}+\mathbf{P})^\top = \mathbf{X}\mathbf{X}^\top + \mathbf{X}\mathbf{P}^\top + \mathbf{P}\mathbf{X}^\top + \mathbf{P}\mathbf{P}^\top \tag{11}$$

Given that $p_i = p_j$ for $t_i = t_j$, the matrices $\mathbf{X}\mathbf{P}^\top$ and $\mathbf{P}\mathbf{X}^\top$ are symmetric and contain identical scalar products where timestamps are equal, implying:

$$\mathbf{X}\mathbf{P}^\top = \mathbf{P}\mathbf{X}^\top \tag{12}$$

Thus, $\mathbf{X}\mathbf{P}^\top + \mathbf{P}\mathbf{X}^\top$ is doubly symmetric and dependent only on the positional encodings, independent of the order of elements sharing the same timestamp.

Analysis of $\mathbf{P}\mathbf{P}^\top$ The term $\mathbf{P}\mathbf{P}^\top$ is invariant under permutations of elements with the same timestamp, as it depends solely on the positional encodings $p(t_i)$, which are identical for elements sharing the same timestamp.

Therefore, the output of the attention mechanism, which depends on A, remains unchanged under permutations of elements x_i with the same timestamp t_j. This demonstrates that the attention mechanism is permutation invariant for elements with the same timestamp when the same positional encoding is applied.

This property is crucial for applications where the order of inputs within the same timestamp does not convey additional information, ensuring that the attention mechanism respects the inherent symmetry of the data.

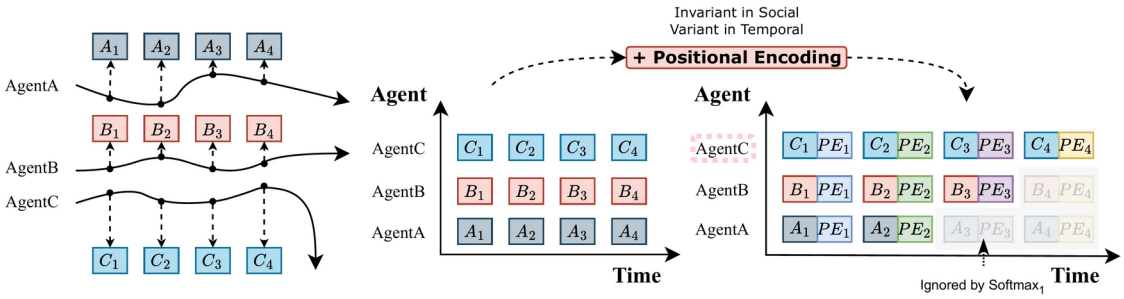

Fig. 2. Overview of Out Way Attention. The improved attention mechanism through Softmax can ignore irrelevant vehicles. As shown in the figure, when the target vehicle C makes a turning behavior, Softmax_1 can ignore the historical trajectories of neighboring agents A and B, which no longer affect C. Gray blocks represent historical and social features that do not need to be considered.

Encoder

We use the same encoder as the Autobots model[39], taking the set of states of each agent over a past observation period $X = (X_1, X_2, \cdots X_{T_{obs}})$ as the input to the encoder, resulting in a three-dimensional tensor with dimensions *K*, *t*, and *M*.

The encoder uses the improved Multi-Head Out Way Attention (OWA) module to process the *temporal* relationships between agents. We first process the time dimension. After adding a sine position encoding to the historical features of each agent, we perform the following operation for each agent seperately:

$$X_{:,m}^{l+1} = \mathrm{OWA}(\mathrm{FFN}(X_{:,m}^{l})) \tag{13}$$

where denotes the layer index, and m specifies a particular column or feature dimension. The colon indicates that we are considering all rows or elements along that dimension, effectively selecting the entire column m.

This equation seems to be a part of a neural network, possibly a transformer or a related architecture, where each layer performs specific transformations on the input data. The specific operations in this equation are as follows:

Where FFN is a feed-forward neural network that can project each element in the set into hidden state space, and FFN itself is not affected by the order of input.

Secondly, to process the spatial interaction between agents, we perform the following operation for the set of *Tobs* observation lengths in turn:

$$X^{l+1} = \mathrm{OWA}(X^{l}) \tag{14}$$

After repeating L times, we obtain a tensor $X^{L} \in \mathbb{R}^{d_K, M, t}$. L here refers the num of layers.

Require: Historical trajectories $\mathbf{X} \in \mathbb{R}^{M \times T_{obs} \times K}$, HD map features $\mathscr{M} \in \mathbb{R}^{N_{poly} \times N_{pts} \times 3}$, Number of proposals $c \in \mathbb{N}$, Prediction horizon $T_{pred} \in \mathbb{N}$
Ensure: Predicted trajectories $\hat{Y} \in \mathbb{R}^{M \times c \times T_{pred} \times 2}$, Mode probabilities $\mathbf{p} \in \mathbb{R}^{c}$

// —— *Initialization* ——
$H^0 \leftarrow \mathrm{LinearProj}(\mathbf{X})$ // *Project to hidden dimension d*
$P_{temp} \leftarrow \mathrm{SinusoidalPE}(T_{obs})$ // *Temporal positional encoding*
$P_{social} \leftarrow \mathrm{ZeroEncoding}(M)$ // *Social permutation encoding*
// —— *Input Processing* ——
for $m = 1$ **to** M **do**
  $H^0_{:,m,:} \leftarrow H^0_{:,m,:} + P_{temp} + P_{social}(m)$ // *Add combined encodings*
end for
// —— *Encoder Processing* ——
for $l = 1$ **to** L_{enc} **do**
  for $m = 1$ **to** M **do**
    $\tilde{H}^l_m \leftarrow \mathrm{LayerNorm}(\mathrm{FFN}(H^{l-1}_{:,m,:}))$ // *Temporal processing*
    $H^l_{:,m,:} \leftarrow \mathrm{OWA}(\tilde{H}^l_m, \tilde{H}^l_m, \tilde{H}^l_m)$ // *Out-Way self-attention*
  end for
  $\bar{H}^l \leftarrow \mathrm{LayerNorm}(H^l)$ // *Social interaction preparation*
  $H^l \leftarrow \mathrm{OWA}(\bar{H}^l, \bar{H}^l, \bar{H}^l)$ // *Agent-agent attention*
end for
// —— *Context Fusion* ——
$C \leftarrow \mathrm{MapEncoder}(\mathscr{M}) \oplus \mathrm{MeanPool}(H^{L_{enc}})$ // *Map-trajectory fusion*
$\{Q_i\}_{i=1}^{c} \leftarrow \mathrm{LearnableParams}(c, d)$ // *Proposal matrices*
// —— *Multimodal Decoding* ——
for $i = 1$ **to** c **do**
  $C_i \leftarrow \mathrm{FFN}(C \oplus Q_i)$ // *Proposal-specific context*
  $O_i \leftarrow \mathrm{Decoder}(C_i, H^{L_{enc}})$ // *OWAD decoding*
  $\hat{Y}_i \leftarrow \mathrm{TrajHead}(O_i)$ // *Trajectory generation*
  $p_i \leftarrow \mathrm{ProbHead}(\mathrm{MaxPool}(O_i))$ // *Mode scoring*
end for
// —— *Final Output* ——
$\hat{Y} \leftarrow \mathrm{TopK}(\{\hat{Y}_i\}, k = 5)$ // *Select best trajectories*
$\mathbf{p} \leftarrow \mathrm{Softmax}([p_1, ..., p_c])$ // *Normalize probabilities*
return $\hat{Y}, \mathbf{p}$

Algorithm 1. PowFormer

Decoder

Future trajectory predictions are inherently fuzzy, and multiple plausible future trajectories can be generated under conditions of identical trajectories. If only one future trajectory is generated, there may be overfitting or even unrealistic trajectories. We abandon the commonly used autoregressive decoding method and train using a method that can learn the proposal matrix.

To more fully capture and mine high-order hidden interaction information, we introduce discrete latent variables *Z*. The introduction of discrete latent variables can also simplify the problem, making the calculation of the likelihood function more accurate and convenient.

$$P(Y_{T_{obs+1}}|X_{T_{obs}},\cdots X_1)=\sum_{Z}P(Y_{T_{obs+1}},Z|X_{T_{obs}},\cdots X_1) \tag{15}$$

We assign a proposal parameter matrix Q_i to each latent variable z_i and train the randomly generated proposal matrix with the aim of outputting a latent target trajectory based on each proposal matrix.

The map information is integrated with the original proposal parameters and jointly used as the input to the decoding module FFN. The context tensor *C* is then expanded in turn according to the time dimension and social dimension. After FFN processing, a tensor $H \in \mathbb{R}^{d_K,M,T}$ is obtained. These proposal matrices will correspond to different future trajectories, enabling the prediction of multi-modal distributions in the trajectory space, thus achieving a one-to-many mapping. At the same time, all proposal matrices can be calculated in parallel, changing the step-by-step calculation method in the autoregressive model and improving the running speed of the model. Each agent is processed by the Out-Way Attention Decoder (OWAD) layer:

$$H_m^{'}=\mathrm{OWAD}(H_m,C_m) \tag{16}$$

The decoding process is repeated *c* times, each time using a different learnable proposal matrix and additional context information, finally outputting a tensor $O \in \mathbb{R}^{d_K,M,T,c}$, which is then processed by a neural network to generate the future trajectory distribution, ultimately achieving the generation of c trajectory suggestions for M agents. Benefiting from the division of the proposal matrix, we address the shortcomings of the implicit coding model that is difficult to train and the limited flexibility of the display coding model, and realize the operational speedup by means of parallel computation.

Finally, the overall process of PowFormer is represented using Algorithm 1.

Experiments

Experimental settings

Datasets

In this section, we describe the datasets used in our experiments, including Argoverse, TrajNet++, and ETH/UCY.

- **Argoverse prediction dataset** Argoverse is the first dataset that includes high-definition maps. The dataset provides the trajectory and location of the ego vehicle and neighboring vehicles, scene metadata, labels and targets, and map data. For each scene, map information is represented as a polyline based on the centerline. The dataset is divided into 5-second intervals, where the trajectory of the target vehicle in the past 2 seconds is given (with a sampling frequency of 10Hz), and the task is to predict the trajectory within the next 3 seconds.
- **TrajNet++ dataset** TrajNet++ is a pedestrian benchmark dataset centered on interaction-based trajectories. It contains a total of 54,513 unique scenes, which are divided into 49,062 training scenes and 5,451 validation scenes. We use the synthetic partition of this dataset, which includes more complex interaction behaviors between agents. The dataset is captured in real road conditions and contains rich information. The complex interactions between different agents cannot be captured using manually set thresholds.
- **ETH/UCY dataset** ETH is a dataset for pedestrian detection. The testing set contains 1,804 images across three video clips. The dataset is captured from a stereo rig mounted on a car, with a resolution of 640 × 480 (bayered) and a framerate of 13–14 FPS.

Metrics We follow the Argoverse benchmark and use minimum average displacement error (minADE), minimum final displacement error (minFDE), and missing rate (MR) as the validation metrics.

- **minADE** The minimum value of several average Euclidean (L2) distance between the predicted trajectories and the actual trajectory.
- **minFDE** The minimum L2 distance between the final predicted trajectory and the actual trajectory among *K* predictions.
- **MR** The ratio of cases where the L2 distance between all *K* predictions and the ground truth at the final time step *T* is greater than 2m.

$$\mathrm{minADE}=\min_{K}\frac{1}{T_{obs}}\sum_{t=1}^{T_{obs}}||Y_t^k-X_t||_2 \tag{17}$$

$$\mathrm{minFDE}=\min_{K}||Y_{T_{pred}}^k-X_{T_{pred}}||_2 \tag{18}$$

$$\mathrm{MR} = \begin{cases} 0, \exists k \in \{1, \cdots, K\} \, ||Y_i^{T_{pred}} - X_i^{T_{pred}}||_2 \leq 2 \\ 1, otherwise \end{cases} \tag{19}$$

Implementation details

We implemented our model utilizing the PyTorch open-source framework and trained it with the Adam optimizer[40], initializing the learning rate at $1e-4$. A comprehensive grid search was conducted, ultimately confirming the selection of this optimizer and learning rate. The specific experimental results are illustrated in Fig. 3.

We followed the model structure settings and objective function of AutoBot[39]. In all datasets, the parameters of the Transformer network are set as follows: the number of Encoder layers is 2, the number of heads in the multi-head attention mechanism is set to 8, and the number of hidden layer units is 384; the number of Decoder layers is 2, and the number of hidden layer units is 384; the hidden dimension of the feed-forward layer is 1536. Our model uses a dropout rate of 0.1 in all parts of the encoder and decoder.

Hardware efficient implementations

We set $Q \in \mathbb{R}^{n\times d}$,$K \in \mathbb{R}^{n\times d}$,$V \in \mathbb{R}^{n\times d}$,$K^\top \in \mathbb{R}^{d\times n}$. In the implementation of the attention mechanism, since Softmax is a fixed coda operator, we add two all-zero vectors to the end of the sequences $K^\top$ and V, i.e., $K^{T'} = (n, n+1)$, let $z = \frac{QK^{T'}}{\sqrt{d}}, i \in (1, n+1)$

$$\mathrm{Softmax}_1(z_i) = \frac{e^{z_i}}{\sum_{c=1}^{n+1} e^{z_c}} = \frac{e^{z_i}}{\sum_{c=1}^{n} e^{z_c} + e^0} = \frac{e^{z_i}}{\sum_{c=1}^{n} e^{z_c} + 1} \tag{20}$$

$$\mathrm{Head_i} = \mathrm{OutWayAttention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{Softmax}_1(\frac{\mathrm{QK}^{\top'}}{\sqrt{d}})V' \tag{21}$$

$$\mathrm{MultiHead(Q,K,V)} = \mathrm{concat}(\mathrm{Head}_1, \mathrm{Head}_2, \cdots, \mathrm{Head_N})W^o \tag{22}$$

When there are a large number of infinitesimal values, 0 is relatively large at this time and is the place where attention is more concentrated. However, under normal circumstances, 0 will not attract extra attention and will not affect the distribution of attention. This scheme provides a zero-exit for the calculation run, achieving the effect of $Softmax_1$.

Interactive attention analysis

In order to more intuitively illustrate the role of the improved attention mechanism, we randomly selected a traffic scenario in the Argoverse dataset. We'll number the cars as shown in Fig. 4. We extract the attention matrix for analyzing the level of attention between different agents. In order to show the effect of the improved attention mechanism more clearly, we extract the same for the pre-improved attention matrix in the temporal dimension and spatial dimension.

Figure 5 shows the distribution of attention in the temporal dimension, Fig. 6 shows the distribution of attention in the social dimension, with both sets of plots the left side is before the improvement of attention and the right side is after the improvement of attention. The axes in Fig. 5 represent the time, and the axes in Fig. 6 represent the number of agent. We can see that the proportion of darker colors in the right side of the picture is higher than in the left side, which explains the fact that our attention mechanism ignores some of the time and agents. For example, in Fig. 6 when considering the influence of vehicles 6 and 7 on vehicle 1, Out-Way attention

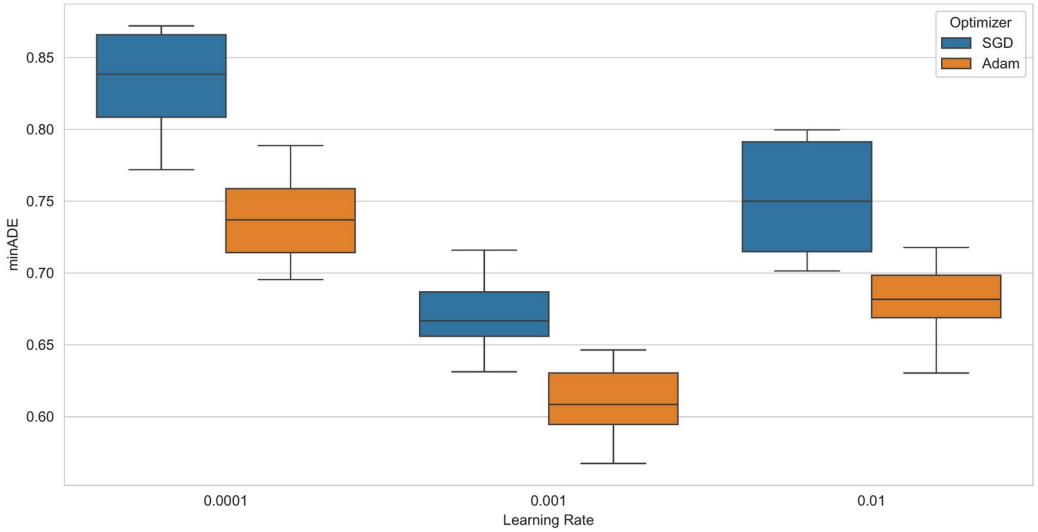

Fig. 3. Grid search of learning rate and optimizer.

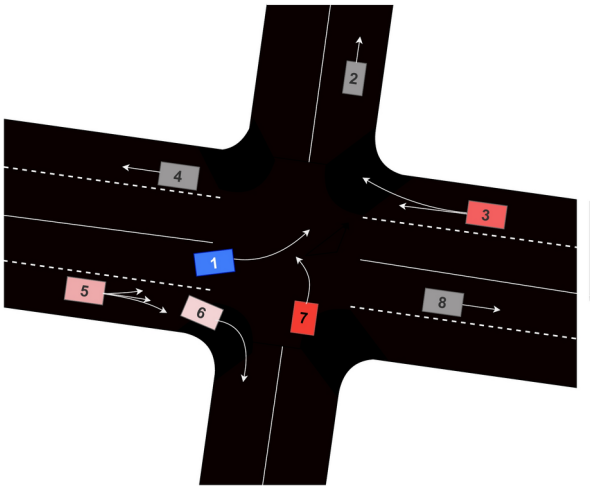

Fig. 4. Vehicle marking chart. This image displays the traffic flow conditions at a relatively complex intersection. The blue color represents the ego agent, the colors of the neighbor agents represent the degree of relevance to the ego agent. Darker colors represent stronger correlations, and grey represents vehicles that can be ignored in predictions. Thanks to our Out Way attention mechanism, we can achieve ignorance for close but irrelevant agents (e.g., between 6 and 1), so that the ego agent's attention is more focused on other agents.

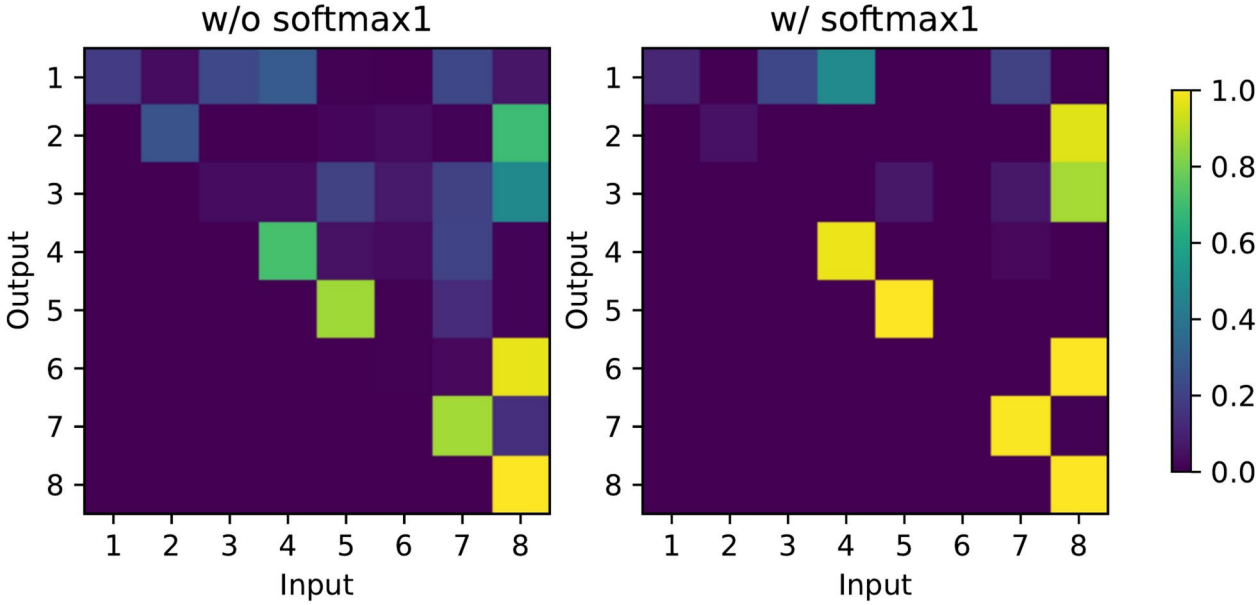

Fig. 5. Comparison of vehicle temporal attention matrices. In the two figures, the use of the $\mathrm{Softmax}_1$ mechanism results in a darker color, indicating that the majority of the attention scores tend to be 0. This shows that our method is capable of ignoring unimportant agents.

can achieve rapid identification and adjustment of attention allocation after the steering behavior of vehicles 6 and 7 to improve prediction.

Through visualize the difference between before and after the improved attention matrix, the effect of Out Way attention on the allocation of attention was verified, and the validity of this effect was demonstrated through the analysis of the results in the later section.

Baselines

We selected some classic, highly relevant, and relatively recent methods as baselines, which are introduced below:

- **DenseTNT** [25]: An endpoint-based trajectory prediction strategy, directly predicting the probability distribution of the endpoint based on TNT;

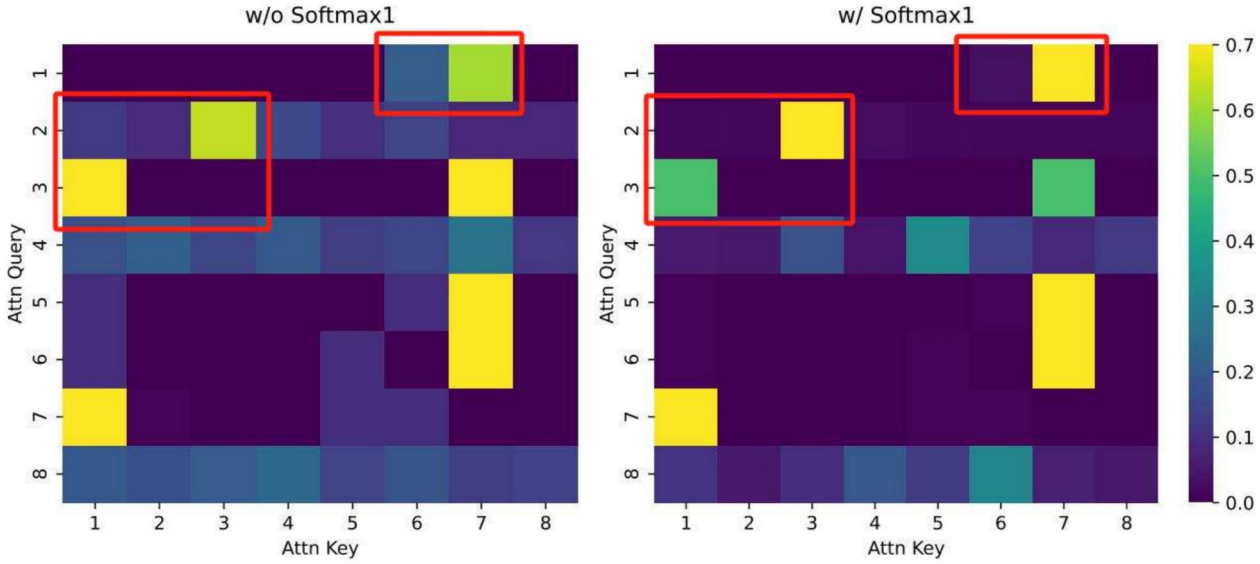

Fig. 6. Comparison of vehicle social attention matrices. As shown in the figure, our $\mathrm{Softmax}_1$ brings the overall attention score closer to 0 (which means a darker background color), thereby ignoring agents that are not important for the current trajectory prediction. Specifically, we can examine the pair of agents (1,6), who turn in opposite directions, hence their mutual attention score should be relatively low. Our $\mathrm{Softmax}_1$ has identified this phenomenon and reflected it in the changes of the attention scores.

- **mmTransformer** [8]: Designs a stacked Transformer network structure, proposing a region-based training strategy;
- **GOHOME** [41]: Uses high-definition maps and sparse projections to generate heat maps, outputting the probability distribution of future locations through the heat map;
- **HiVT-128** [10]: Proposes a hierarchical vector Transformer model, a multi-agent motion prediction algorithm capturing local and global interactions;
- **Social-BiGAT** [42]: Proposes a graph-based generative adversarial network for generating realistic, multimodal trajectory predictions for multiple pedestrians in a scene.
- **AgentFormer** [43]: This model generates sequence representations by flattening the features of multi-agent trajectories across time and agents.
- **AutoBots** [39]: Proposes a multimodal latent sequence encoding-decoding mode based on Transformer.
- **PPT** [19]: Proposed a novel Progressive Pretext Task framework, which enhances the model's ability to capture short-term dynamics and long-term dependencies through gradual training, thereby improving its performance on the final overall trajectory prediction task.

For different datasets, we used different methods as baselines. This is because some methods are specifically designed for vehicle/pedestrian trajectory prediction. On the other hand, we lack sufficient computational resources to train and fine-tune these methods on different datasets.

Performance results

In this section, we present the results on the self-driving dataset Argoverse and the pedestrian dataset Trajnet++. Figure 7 compares the trends of minADE and minFDE on the Argoverse dataset and Trajnet++ dataset respectively when the output modal number is 2, 4, 6, 8, 10.

Trend analysis of output modes Figure 7 shows the trend of the evaluation metrics as the number of predicted trajectories changes. As the number of prediction modes increases, both "best minADE" and "best minFDE" decrease, indicating that the model's prediction accuracy improves with more modes. This demonstrates that our model exhibits good prediction performance as the number of output modes increases.

Performance on the Argoverse dataset Table 1 shows the comparison of our method, PowFormer, with the baseline models on the Argoverse dataset when $c = 6$. PowFormer achieves state-of-the-art performance by significantly reducing reliance on artificial rules and leveraging the enhanced Out Way Attention (OWA) mechanism. Specifically, PowFormer achieves the best result in minADE (0.65) and surpasses Autobots in minFDE by 19%. The improvement in minADE can be attributed to the OWA mechanism's ability to better capture fine-grained interactions and adapt to diverse agent behaviors.

While the improvement in minFDE compared to Autobots is less pronounced, PowFormer still achieves a competitive value of 1.10, demonstrating its robustness in predicting final destinations. Furthermore, PowFormer achieves the lowest Miss Rate (MR) of 0.10, outperforming HiVT-128 (0.09) and other baselines. This indicates that our method effectively reduces prediction errors in complex multi-agent scenarios, ensuring reliable trajectory forecasts.

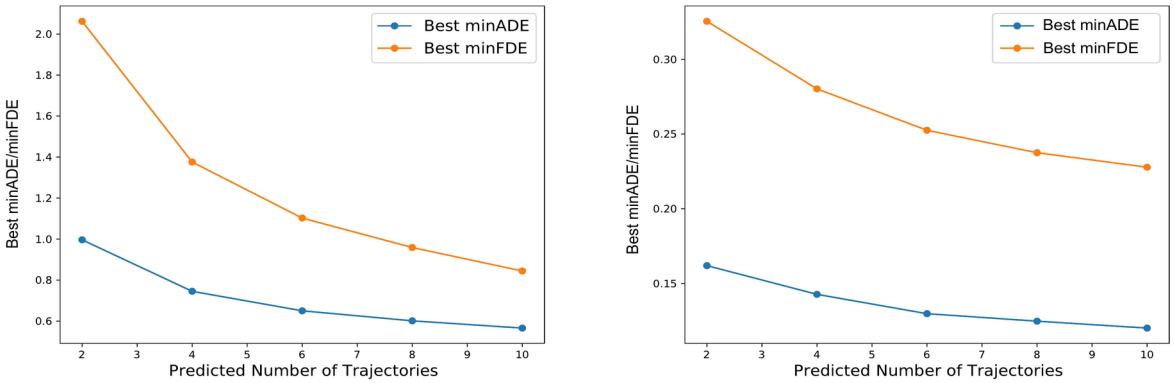

Fig. 7. Best minADE/minFDE with different number of modes on Argoverse (left) and Trajnet++ (right) datasets.

Method	minADE	minFDE	MR
DenseTNT	0.88	1.28	0.11
mmTransformer	0.83	1.29	0.15
GOHOME	0.94	1.45	0.11
HiVT-128	0.77	1.69	**0.09**
AutoBots	0.73	1.10	0.11
PowFormer (Ours)	**0.65**	**1.10**	0.10

Table 1. Comparison of our method and the baseline in terms of performance on the Argoverse dataset.

Method	minADE	minFDE
Social-BiGAT	0.26	0.52
Social LSTM	0.24	0.49
AutoBots	0.13	0.24
PowFormer (Ours)	**0.12**	**0.22**

Table 2. Performance on the Trajnet++.

Performance on the ETH/UCY dataset To further evaluate PowFormer's ability to handle pedestrian trajectory prediction in dense and interactive environments, we test it on the ETH/UCY dataset, as shown in Table 3. PowFormer achieves competitive performance with a minADE of 0.21 and a minFDE of 0.30. Notably, PowFormer outperforms AgentFormer in minFDE by 23.1% and achieves comparable results with PPT, which has the best minADE (0.20). This improvement highlights PowFormer's ability to accurately model human social interactions and predict long-term trajectories in crowded environments.

Compared to Social BiGAT, which heavily relies on graph-based modeling, PowFormer demonstrates a significant improvement of 56.3% in minADE and 64.3% in minFDE. These results validate the effectiveness of the OWA mechanism in capturing both global and local interaction patterns, enabling PowFormer to generalize well to pedestrian datasets with diverse motion patterns.

Generalization on the TrajNet++ dataset To assess PowFormer's generalization capability, we evaluate it on the synthetic partition of the TrajNet++ dataset under the single-modal setting ($c = 1$). Table 2 shows that PowFormer achieves the best performance with a minADE of 0.12 and a minFDE of 0.22, surpassing Autobots by 7.7% in minADE and 8.3% in minFDE. Moreover, PowFormer outperforms Social LSTM by 45.8% in minADE and 48.9% in minFDE, demonstrating its ability to handle complex multi-agent environments and accurately predict trajectories.

These results indicate that PowFormer effectively captures the interaction relationships between agents and generalizes well to full-domain scenarios, even in synthetic datasets with diverse motion patterns and noise. The superior performance on TrajNet++ underscores the robustness of our OWA-based architecture in handling both real-world and synthetic datasets.

Comparison across datasets As shown in Tables 1, 2, and 3, PowFormer consistently outperforms state-of-the-art methods across diverse datasets, including Argoverse, TrajNet++, and ETH/UCY. The key to this improvement lies in the enhanced OWA mechanism, which effectively ignores irrelevant factors and focuses on meaningful interactions. This enables PowFormer to accurately capture potential trajectories while balancing global and local information.

Method	minADE	minFDE
Social BiGAT	0.48	0.84
AgentFormer	0.23	0.39
PPT	**0.20**	0.31
PowFormer (Ours)	0.21	**0.30**

Table 3. Performance on ETH/UCY.

Method	Model component			Evaluation metric	
	Encoder–Decoder	PPM	OWA	minADE	minFDE
Transformer	✓			0.21	0.56
Transformer w/o Out-Way	✓	✓		0.14	0.23
Transformer w/o Proposal	✓		✓	0.15	0.49
PowFormer	✓	✓	✓	**0.12**	**0.22**

Table 4. Ablation study of trajectory prediction for each sub-module.

Num. Heads	minADE	minFDE
1	0.69	1.18
2	0.67	1.14
4	0.66	1.12
8	0.65	1.11
12	0.65	1.10

Table 5. The impact of attention heads on model performance. We set same number of attention heads for both encoder and decoder.

For example, in the Argoverse dataset, PowFormer achieves a 19% improvement in minFDE compared to Autobots, demonstrating its strength in predicting final destinations. In the ETH/UCY dataset, PowFormer achieves comparable results with PPT in minADE while surpassing AgentFormer in minFDE by 23.1%. On the TrajNet++ dataset, PowFormer demonstrates its generalization ability by achieving the best performance in both minADE and minFDE, outperforming Social LSTM by a large margin.

In summary, PowFormer achieves state-of-the-art performance across Argoverse, ETH/UCY, and TrajNet++ datasets. The improved OWA mechanism plays a pivotal role in enhancing prediction accuracy and generalization ability, proving the effectiveness of our approach in diverse and challenging scenarios.

Ablation studies

To verify the effectiveness of the proposal matrix as well as the Out Way attention mechanism, we conduct an ablation study. We divided the PowFormer into different parts and performed experiments to quantify its role by reducing one or more of them. The most basic Transformer encoding and decoding structures are included in these experiments, and the same parameters are used to train all models. We first remove one module alternately to study the effect of different modules on the prediction performance, and finally perform all removals. We compare PowFormer to three ablated models: (1) The sub-models tested include the base Transformer encoder–decoder model, (2) the proposal parameters model (PPM) that includes a Transformer encoder–decoder architecture and proposal parameters, (3) the Out Way attention model (OWA) that includes a Transformer encoder–decoder architecture and improved attention mechanisms. Our experiments were performed on the Trajnet++ dataset.

As can be seen from Table 4, the results of the ablation experiments indicate that utilizing the same parameters with the same dataset, the prediction accuracy decreases to varying degrees when different sub-models are removed in the PowFormer. Obviously, sub-models lacking proposal matrices tend to fall into the misconception of modal homogeneity and have lower prediction accuracies. Sub-models that do not include Out Way attention are less capable of capturing global attention mechanisms. Therefore each prediction module is critical to the prediction task.

The proposed model employs enhanced multi-head attention modules extensively. It is imperative to investigate how the number of attention heads in both encoder and decoder components influences the model's predictive accuracy. Denoting the multi-head attention mechanisms in the encoder and decoder as h-encoder and h-decoder respectively, the performance variations under different head configurations were evaluated using minADE and minFDE metrics. The comparative results across various combinations of attention heads in both components are illustrated in the Table 5.

Conclusion

In this paper, we propose the Powformer prediction method, which can achieve the prediction of multiple potential trajectories for agents in the future. We introduce a new attention mechanism, which we call Out Way Attention. This attention can fully capture and reasonably utilize scene information, generalize the original local scene, and avoid falling into local optimum. PowFormer is learned and evaluated using the Argoverse dataset and Trajnet++ dataset, which contains trajectories of real-world vehicles and other agents. Through the in-depth study of the attention mechanism, a more adequate capture of local and global features can be realized. Combined with the latent sequence Transformer encoding-decoding mode for multi-modal trajectory prediction tasks, compared with the original Autobots model, our improved attention mechanism model has improved results. More importantly, our improved attention mechanism provides ideas that can be applied in other studies with generalizability.

Data availability

The datasets underpinning the findings of this research are publicly accessible. The Argoverse dataset can be accessed at https://www.argoverse.org/, the Trajnet++ dataset is available through https://www.epfl.ch/labs/vita/research/prediction/socially-aware-human-trajectory-forecasting/trajnet/, and the ETH/UCY dataset is available through https://icu.ee.ethz.ch/research/datsets.html.

Received: 20 January 2025; Accepted: 3 April 2025
Published online: 19 April 2025

References

1. Mahajan, V., Katrakazas, C. & Antoniou, C. Prediction of lane-changing maneuvers with automatic labeling and deep learning. *Transp. Res. Rec.* **2674**, 336–347 (2020).
2. Rathore, P., Kumar, D., Rajasegarar, S., Palaniswami, M. & Bezdek, J. C. A scalable framework for trajectory prediction. *IEEE Trans. Intell. Transp. Syst.* **20**, 3860–3874. https://doi.org/10.1109/TITS.2019.2899179 (2019).
3. Houenou, A., Bonnifait, P., Cherfaoui, V. & Yao, W. Vehicle trajectory prediction based on motion model and maneuver recognition. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems* 4363–4369. https://doi.org/10.1109/IROS.2013.6696982 (2013).
4. Leon, F. & Gavrilescu, M. A review of tracking and trajectory prediction methods for autonomous driving. *Mathematics* **9**, 660. https://doi.org/10.3390/math9060660 (2021).
5. Cong, P. *et al.* Dacr-amtp: Adaptive multi-modal vehicle trajectory prediction for dynamic drivable areas based on collision risk. *IEEE Trans. Intelli. Vehicles* 1–22. https://doi.org/10.1109/TIV.2023.3321656 (2023).
6. Huang, R., Xue, H., Pagnucco, M., Salim, F. D. & Song, Y. Multimodal trajectory prediction: A survey. arXiv arXiv:abs/2302.10463 (2023).
7. Huang, Z., Liu, H. & Lv, C. Gameformer: Game-theoretic modeling and learning of transformer-based interactive prediction and planning for autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* 3880–3890. https://doi.org/10.1109/ICCV51070.2023.00361 (2023).
8. Liu, Y., Zhang, J., Fang, L., Jiang, Q. & Zhou, B. Multimodal motion prediction with stacked transformers. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 7573–7582. https://doi.org/10.1109/CVPR46437.2021.00749 (2021).
9. Zhu, Y., Luan, D. & Shen, S. Biff: Bi-level future fusion with polyline-based coordinate for interactive trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* 8260–8271 (2023).
10. Zhou, Z., Ye, L., Wang, J., Wu, K. & Lu, K. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 8813–8823. https://doi.org/10.1109/CVPR52688.2022.00862 (2022).
11. Zheng, F. *et al.* Unlimited neighborhood interaction for heterogeneous trajectory prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* 13148–13157. https://doi.org/10.1109/ICCV48922.2021.01292 (2021).
12. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S. & Alahi, A. Social gan: Socially acceptable trajectories with generative adversarial networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* 2255–2264. https://doi.org/10.1109/CVPR.2018.00240 (2018).
13. Zhao, H. *et al.* Tnt: Target-driven trajectory prediction. arXiv:2008.08294 (2020).
14. Chai, Y., Sapp, B., Bansal, M. & Anguelov, D. Multipath: Multiple probabilistic anchor trajectory hypotheses for behavior prediction (2019). arXiv:1910.05449.
15. Yuan, Y., Weng, X., Ou, Y. & Kitani, K. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9793–9803, https://doi.org/10.1109/ICCV48922.2021.00967 (2021).
16. Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inform. Process. Syst.* **30** (2017).
17. Mignone, P., Corizzo, R. & Ceci, M. Distributed and explainable Ghsom for anomaly detection in sensor networks. *Mach. Learn.* **113**, 4445–4486 (2024).
18. Qaddoura, R., Younes, M. B. & Boukerche, A. Towards optimal tuned machine learning techniques based vehicular traffic prediction for real roads scenarios. *Ad Hoc Netw.* **161**, 103508 (2024).
19. Lin, X., Liang, T., Lai, J. & Hu, J.-F. Progressive pretext task learning for human trajectory prediction (2024). arXiv:2407.11588.
20. Chang, M.-F. *et al.* Argoverse: 3d tracking and forecasting with rich maps. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 8740–8749. https://doi.org/10.1109/CVPR.2019.00895 (2019).
21. Kothari, P., Kreiss, S. & Alahi, A. Human trajectory forecasting in crowds: A deep learning perspective. *IEEE Trans. Intell. Transp. Syst.* **23**, 7386–7400. https://doi.org/10.1109/TITS.2021.3069362 (2022).
22. Liang, R., Li, Y., Zhou, J. & Li, X. Stglow: A flow-based generative framework with dual-graphormer for pedestrian trajectory prediction. *IEEE Trans. Neural Netw. Learn. Syst.* 1–14. https://doi.org/10.1109/TNNLS.2023.3294998 (2023).
23. Bhattacharyya, A., Straehle, C. N., Fritz, M. & Schiele, B. Haar wavelet based block autoregressive flows for trajectories. *Pattern Recogn.* **12544**, 275–288 (2020).
24. Phan-Minh, T., Grigore, E. C., Boulton, F. A., Beijbom, O. & Wolff, E. M. Covernet: Multimodal behavior prediction using trajectory sets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).
25. Gu, J., Sun, C. & Zhao, H. Densetnt: End-to-end trajectory prediction from dense goal sets (2021). arXiv:2108.09640.
26. Wu, D. & Wu, Y. air^2 for interaction prediction (2021). arXiv:2111.08184.
27. Alahi, A. *et al.* Social lstm: Human trajectory prediction in crowded spaces. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 961–971. https://doi.org/10.1109/CVPR.2016.110 (2016).

28. Deo, N. & Trivedi, M. M. Convolutional social pooling for vehicle trajectory prediction. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* 1549–15498. https://doi.org/10.1109/CVPRW.2018.00196 (2018).
29. Mohamed, A., Qian, K., Elhoseiny, M. & Claudel, C. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 14412–14420. https://doi.org/10.1109/CVPR42600.2020.01443 (2020).
30. Vemula, A., Muelling, K. & Oh, J. Social attention: Modeling attention in human crowds. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* 4601–4607. https://doi.org/10.1109/ICRA.2018.8460504 (2018).
31. Wen, F., Li, M. & Wang, R. Social transformer: A pedestrian trajectory prediction method based on social feature processing using transformer. In *2022 International Joint Conference on Neural Networks (IJCNN)* 1–7. https://doi.org/10.1109/IJCNN55064.2022.9891949 (2022).
32. Cai, Y. et al. Environment-attention network for vehicle trajectory prediction. *IEEE Trans. Veh. Technol.* **70**, 11216–11227. https://doi.org/10.1109/TVT.2021.3111227 (2021).
33. Guo, H. et al. Vehicle trajectory prediction method coupled with ego vehicle motion trend under dual attention mechanism. *IEEE Trans. Instrum. Meas.* **71**, 1–16. https://doi.org/10.1109/TIM.2022.3163136 (2022).
34. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778. https://doi.org/10.1109/CVPR.2016.90 (2016).
35. Gilles, T., Sabatini, S., Tsishkou, D. V., Stanciulescu, B. & Moutarde, F. Thomas: Trajectory heatmap output with learned multi-agent sampling. arXiv arXiv:abs/2110.06607 (2021).
36. Zhang, K. & Li, L. Explainable multimodal trajectory prediction using attention models. *Transp. Res. C: Emerg. Technol.* **143**, 103829. https://doi.org/10.1016/j.trc.2022.103829 (2022).
37. Ngiam, J. *et al.* Scene transformer: A unified architecture for predicting future trajectories of multiple agents. In *International Conference on Learning Representations* (2022).
38. Dash, S., Bagchi, S., Mihindukulasooriya, N. & Gliozzo, A. Permutation invariant strategy using transformer encoders for table understanding. In Carpuat, M., de Marneffe, M.-C. & Meza Ruiz, I. V. (eds.) *Findings of the Association for Computational Linguistics: NAACL 2022* 788–800. https://doi.org/10.18653/v1/2022.findings-naacl.59 (Association for Computational Linguistics, aSeattle, United States, 2022).
39. Girgis, R. *et al.* Latent variable sequential set transformers for joint multi-agent motion prediction. arXiv:2104.00563 (2022).
40. Kingma, D. P. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).
41. Gilles, T., Sabatini, S., Tsishkou, D., Stanciulescu, B. & Moutarde, F. Gohome: Graph-oriented heatmap output for future motion estimation. 9107–9114, https://doi.org/10.1109/ICRA46639.2022.9812253 (2022).
42. Kosaraju, V. *et al.* Social-bigat: multimodal trajectory forecasting using bicycle-gan and graph attention networks. In Wallach, H. *et al.* (eds.) *Advances in Neural Information Processing Systems 32 (NIPS 2019)*, vol. 32 of *Advances in Neural Information Processing Systems* (Neural Information Processing Systems (NIPS), 2019). Funding Information: The research reported in this publication was supported by funding from the TRI gift, ONR (1165419-10-TDAUZ), Nvidia, and Samsung. Publisher Copyright: © 2019 Neural information processing systems foundation. All rights reserved. Copyright: Copyright 2020 Elsevier B.V., All rights reserved.; Advances in Neural Information Processing Systems 2019, NIPS 2019 ; Conference date: 08-12-2019 Through 14-12-2019.
43. Yuan, Y., Weng, X., Ou, Y. & Kitani, K. M. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 9813–9823 (2021).

Acknowledgements
We would like to thank Jitai Hao for valuable support and contributions to this research. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

Author contributions
W.X. designed the study and drafted the manuscript. R.L. performed the statistical analysis. X.D. contributed to data collection and interpretation. B.L. assisted in manuscript preparation and critical revision. L.X. supervised the research and gave final approval for the manuscript. All authors read and approved the final manuscript.

Funding
This work was supported by the Shandong Provincial Social Science Planning Research Project (Grant No. 18CCXJ25), Qingdao Social Science Planning Research Project (Grant No. QDSKL1801134), and the Shandong Provincial Natural Science Foundation Youth Project (Grant No. ZR2023QG039).

Declarations

Consent for publication
All authors consent to the publication of this manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to L.X.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.