Towards Human-like Trajectory Prediction for Autonomous Driving: A Cognitive-inspired Lightweight Model

Haicheng Liao[a], Yongkang Li[b], Zhenning Li[a*],
Chengyue Wang[a], Zilin Bian[c], Ziyuan Pu[d], Jia Hu[e], Zhiyong Cui[f]

[a]University of Macau, [b]UESTC, [c]New York University, [d]Southeast University,
[e]Tongji University, [f]Beihang University
*Corresponding author, Email: zhenningli@um.edu.mo

Extended abstract submitted for presentation at the Conference in Emerging Technologies in Transportation Systems (TRC-30)
September 02-03, 2024, Crete, Greece

April 15, 2024

Keywords: Autonomous Driving, Trajectory Prediction, Knowledge Distillation, Human-like Prediction

1 INTRODUCTION

The emergence of autonomous vehicles (AVs) introduces engineering and cognitive challenges, emphasizing the need for trajectory prediction that encompasses understanding of external environments and internal decision-making mechanisms (Kolekar *et al.*, 2020). The Human-Like Trajectory Prediction (HLTP++) model, inspired by the cognitive processes of human drivers, seeks to address these challenges. It integrates visual processing and decision-making functionalities through a "teacher" model, which employs a neural network with visual pooling reflecting the brain's capability to filter and process visual stimuli. Simultaneously, the "student" model utilizes a novel Spike Neural Network, FA-SNN, to replicate the decision-making efficacy of the prefrontal and parietal cortex (Louie, 2018, Geisslinger *et al.*, 2022). HLTP++ aims to emulate the brain's information handling and decision-making capacity, focusing on adaptability, flexibility, and accuracy, thus contributing significantly to the advancement of AV technology (Dong *et al.*, 2023). Overall, the contributions of HLTP++ are multifaceted:

(1) The HLTP++ emulates human memory and decision-making for traffic trajectory prediction, featuring a novel visual pooling for dynamic observation adjustment across agents and scenes. Additionally, we introduce Fourier Adaptive Spike Neural Network (FA-SNN) for handling incomplete traffic data, inspired by neuronal pulse propagation in human brain.

(2) The HLTP++ introduces a heterogeneous teacher-student framework with Knowledge Distillation Modulation (KDM) for multi-level trajectory prediction tasks. This method dynamically balances loss function ratios, enhancing model training in complex scenarios.

(3) Benchmarking HLTP++ against NGSIM, HighD, and MoCAD datasets reveals it significantly surpasses top baselines, showcasing enhanced robustness and accuracy across diverse traffic conditions, including highways and urban settings. Its performance remains notable under limited input and missing data scenarios.

2 METHODOLOGY

2.1 Problem Formulation

This study focuses on predicting the trajectory of a target vehicle amidst both autonomous vehicles (AVs) and human-driven vehicles, tracking the state of each traffic agent at time t as $p_t^i = (x_t^i, y_t^i)$. Using trajectory data from time $[1, T_{obs}]$ for the target and observed traffic agents,

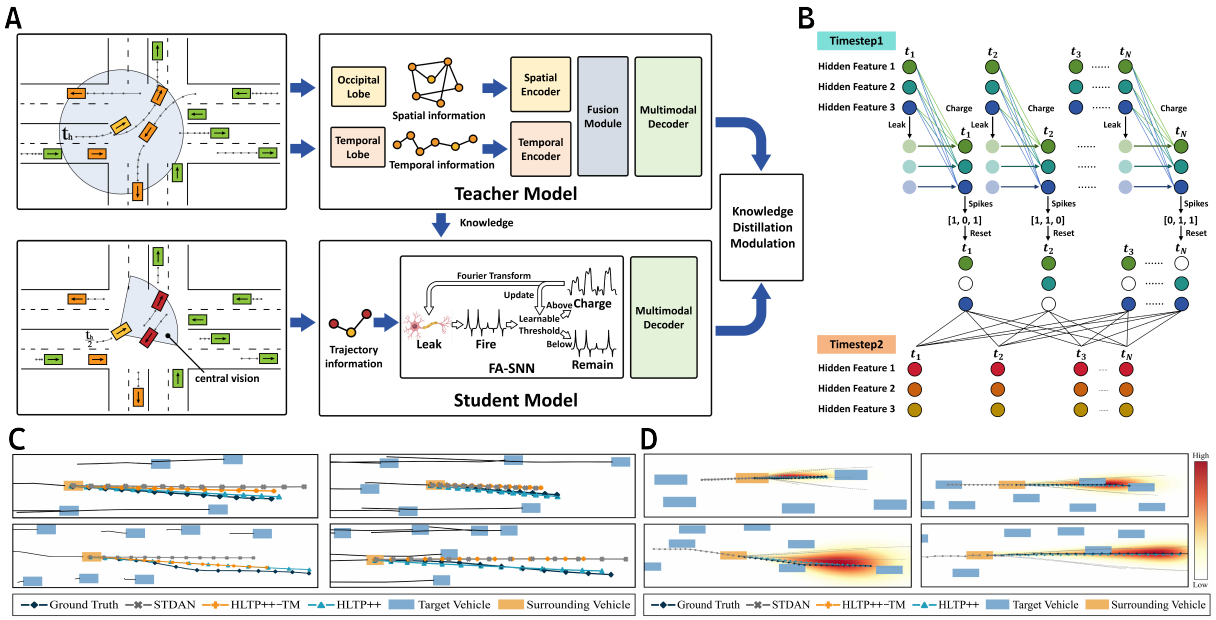

Figure 1 – *Illustration of the HLTP++ structure (A), the FA-SNN structure(B), visualization of the trajectory prediction (C,D).*

denoted $\boldsymbol{X} = \{p_t^{0:n}\}_{t=1}^{T_{obs}}$, the goal is to forecast the future trajectory of the target vehicle with its probabilistic distribution $P(\boldsymbol{Y}|\boldsymbol{X})$, where $\boldsymbol{Y} = \{\boldsymbol{y}_t^0\}_{t=T_{obs}+1}^{T_f} \in \mathbb{R}^{T_f \times 2}$ represents the predicted positions from $T_{obs}+1$ to T_f. Each predicted point $\boldsymbol{y}_t^0$ includes potential trajectories and their likelihood $c_t^{0,i}$, ensuring the sum of likelihoods for all possible maneuvers equals one.

2.2 The Teacher Model

Temporal Encoder. In driving, the human brain optimizes decision-making by prioritizing information, essential for managing its processing capacity and minimizing cognitive load by focusing on relevant features. Our model adopts an LSTM layer for temporal data processing, enhanced with a multi-head attention mechanism for efficient attention distribution.

Spatial Encoder. To replicate human drivers' peripheral monitoring, especially during maneuvers, we introduce the Spatial Encoder. It processes time-segmented matrices $\mathcal{M} \in \mathbb{R}^{(n+1)\times \frac{T_{obs}}{4} \times 2}$ using convolutional layers, batch normalization and dropout to extract features. Then the Graph Attention Networks and ELU activation enhance the spatial features.

Fusion Module. The combined outputs of the Spatial Encoder and the Temporal Encoder are fused and then fed into the iTransformer architecture to generate fused features. Furthermore, we use the disparities between temporal and spatial features to generate a loss, denoted as L_{st}, which serves as one of the loss functions for training the "teacher" model.

Multimodal Decoder. The decoder of the teacher model, based on a Gaussian Mixture Model (GMM), accounts for uncertainty by evaluating multiple possible maneuvers and their probabilities. This multimodal structure not only provides different predictions, but also quantifies their confidence levels, which supports decision-making amidst unpredictability in prediction.

2.3 The Student Model

Visual Pooling Mechanism. We use a novel pooling mechanism with an adaptive visual sector for data preprocessing. This sector is based on the fact that driver's visual area is influenced by speed, which narrows at higher speeds for focused attention and widens at lower speeds for broader awareness. This makes the driver focus more on the central visual field.

FA-SNN. The FA-SNN is an enhanced version of the traditional SNN model, which mimics the neural transmission of the brain. The approach is based on the idea that neurons in SNN should

adapt to different scenarios, which is mainly reflected in the adjustment of the threshold. The FA-SNN involves four essential processes: Charging, Leakage, Firing and Back propagation.
(1) Charging Process. The current neuron is charged by aggregating input spike sequences from previous neurons through varying weights at discrete features.
(2) Leakage Process. Neurons experience leakage due to voltage differences in their surroundings. The internal voltage V of the neuron tends towards an equilibrium voltage U over time t, adhering to the differential equation $U - V = -\tau \frac{dV}{dt}$, where τ denote the leakage decay rate. After solving the above equation, we can obtain $V(t_n) = U - Ce^{-\frac{t_n}{\tau}}$, where C is a constant. This allows us to calculate the voltage at the next moment t_{n+1}: $V(t_{n+1}) = V(t_n + dt) = e^{-\frac{dt}{\tau}}(V(t) - U) + U$
(3) Firing Process. The firing process is activated based on the spike magnitude through an activation function. Given a learnable threshold U_0, the voltage $V'(t_{n+1})$ can be defined as:

$$V'(t_{n+1}) = \begin{cases} F(V(t_{n+1})) - U_0, & V(t_{n+1}) > U_0 \\ F(V(t_{n+1})), & V(t_{n+1}) \leq U_0 \end{cases} \tag{1}$$

where, F is a Fourier Transform.
(4) Back propagation. Due to the discontinuity of the activation function used during SNN firing, conventional chain-rule differentiation is infeasible. To circumvent this, the gradient G is redefined, factoring in the spike threshold and introducing parameters like the absolute width w_a, gradient width w_g and gradient scale s:

$$G(V(t_{n+1})) = \frac{s}{w_a} \times \exp\left(-\frac{|V(t_{n+1}) - U_0|}{w_a}\right) \tag{2}$$

where $w_a = U_0 \cdot w_g$, and $w_g = 0.5$, $s = 1.0$.

2.4 Training

Teacher Training. For the teacher model, we use 3 seconds of observed trajectory for input and predicting a 5-second future trajectory. The loss function of the teacher model is: $\mathcal{L}^{tea} = \mathcal{L}^{tea}_{traj} + \mathcal{L}^{tea}_{man} + \mathcal{L}_{st}$, where $\mathcal{L}^{tea}_{traj} = \sum_t^{T_f} \sum_c^C \mathcal{L}^N(P^{tea}_{pred}, P_{gt})$, $\mathcal{L}^{tea}_{man} = \sum_t^{T_f} \sum_c^C \mathcal{L}^M(M^{tea}_{pred}, M_{gt})$, L_{st} is the temporal-spacial loss from iTransformer. $\mathcal{L}^N$ represents the Negative Log-Likelihood (NLL) loss, $\mathcal{L}^M$ represents the Mean Squared Error (MSE) loss. P^{tea}_{pred} and P_{gt} representing the teacher model's predicted trajectory coordinates and the ground truth coordinates, M^{tea}_{pred} and M^{tea}_{gt} representing the predicted maneuvers and the ground truth maneuvers.
Student Training. The student model is trained to predict 5-second future trajectories with fewer input observations. Similar to the teacher model, the student model has its own loss function $\mathcal{L}^{stu}$ formulated as: $\mathcal{L}^{stu} = \mathcal{L}^{stu}_{traj} + \mathcal{L}^{stu}_{man} = \sum_t^{T_f} \sum_c^C \left(\mathcal{L}^N(P^{stu}_{pred}, P_{gt}) + \mathcal{L}^M(M^{stu}_{pred}, M_{gt})\right)$, where P^{stu}_{pred} and M^{stu}_{pred} represent the predicted 2D coordinates and maneuvers of the student model. Moreover, we apply the MSE loss to measure the disparity between the outputs of the teacher and the student model: $\mathcal{L}^{dis} = \mathcal{L}^{dis}_{traj} + \mathcal{L}^{dis}_{man} = \sum_t^{T_f} \sum_c^C \left(\mathcal{L}^M(P^{stu}_{pred}, P^{tea}_{pred}) + \mathcal{L}^M(M^{stu}_{pred}, M^{tea}_{pred})\right)$. Hence, the total loss function of the "student" model is formulated as $\mathcal{L} = \mathcal{L}^{stu} + \mathcal{L}^{dis}$.
Knowledge Distillation Modulation. We propose a method for tuning multiple tasks that evaluates the importance of different loss functions and automatically adjusts the weights for efficient training based on (Kendall *et al.*, 2018). The tuning loss function is formulated as:

$$\mathcal{L}(W, \sigma_t, \sigma_m, \sigma_s, \sigma_d) = \frac{1}{2\sigma_s^2}\left(\frac{1}{2\sigma_t^2}\mathcal{L}^{stu}_{traj} + \frac{1}{2\sigma_m^2}\mathcal{L}^{stu}_{man}\right) + \frac{1}{2\sigma_d^2}\left(\frac{1}{2\sigma_t^2}\mathcal{L}^{dis}_{traj} + \frac{1}{2\sigma_m^2}\mathcal{L}^{dis}_{man}\right) + F(\sigma_t, \sigma_m, \sigma_s, \sigma_d), \tag{3}$$

where F equals to $log\sigma_t\sigma_m(\frac{1}{2\sigma_s^2} + \frac{1}{2\sigma_d^2}) + log\sigma_s\sigma_d$. To ensure uniformity in the derived equations from different level-segmenting approach, we modify F as $F = log(\sigma_t\sigma_m\sigma_s\sigma_d)$.

3 EXPERIMENT

Our comprehensive evaluation demonstrates HLTP++'s superior performance compared to SOTA baselines, as detailed in Table 2 B-D. The Human-Like Trajectory Prediction model (HLTP++) surpasses existing SOTA models across three datasets, particularly excelling in long-distance

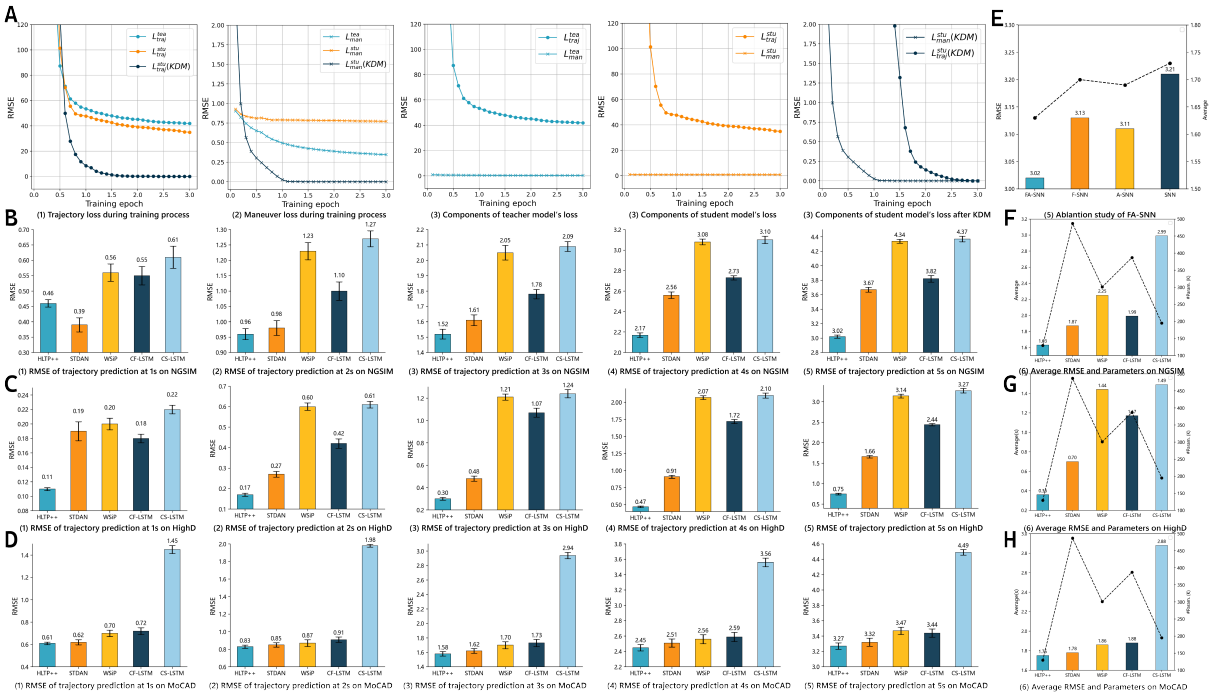

Figure 2 – *Visualization of the experiments' result on NGSIM, HighD, and MoCAD.*

trajectory forecasting. Furthermore, as depicted in Table 2 F-H, our model boasts the smallest parameter count, with the student model having only 129.6K parameters. This represents a 56.91% reduction compared to the previously smallest model, WSiP, while achieving superior performance. Additionally, ablation studies on the FA-SNN, as shown in Table 2 E, underscore the importance of utilizing Fast Fourier Transform and adaptive thresholds. Moreover, Table 2 A reveals that before applying Knowledge Distillation Modulation (KDM), there was a significant discrepancy between different loss functions. However, post-KDM application, the disparity between the loss functions narrows to a similar magnitude, and the overall loss decreases. This indicates the efficacy of KDM in training models with a multitude of loss functions.

4 DISCUSSION

This study introduces HLTP++, a novel trajectory prediction model for autonomous vehicles (AVs), overcoming prior models' limitations with a special knowledge distillation framework. It offers a lightweight, efficient solution that retains accuracy through human-like observation and prediction capabilities. Empirical results demonstrate HLTP++'s superiority in complex traffic environments, achieving state-of-the-art (SOTA) performance. Future efforts will focus on developing models that emulating the complexity of the human brain system, using multimodal data for improved processing and forecasting.

References

Dong, Jiqian, Chen, Sikai, Miralinaghi, Mohammad, Chen, Tiantian, Li, Pei, & Labi, Samuel. 2023. Why did the AI make that decision? Towards an explainable artificial intelligence (XAI) for autonomous driving systems. *Transportation research part C: emerging technologies*, **156**, 104358.

Geisslinger, Maximilian, Poszler, Franziska, & Lienkamp, Markus. 2022. An ethical trajectory planning algorithm for autonomous vehicles. *Nature Machine Intelligence*, **5**, 137–144.

Kendall, Alex, Gal, Yarin, & Cipolla, Roberto. 2018. *Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics.*

Kolekar, Sarvesh, de Winter, Joost, & Abbink, David. 2020. Human-like driving behaviour emerges from a risk-based driver model. *Nature communications*, **11**(1), 1–13.

Louie, Jennifer. 2018. Working memory capacity and executive attention as predictors of distracted driving.