A Chain-of-Thoughts Enhanced LLM-Instructed Trajectory Prediction Framework for Autonomous Driving

Haicheng Liao[a], Hanlin Kong[b], Junxian Yang[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, Large Language Model, Chain-of-Thoughts Reasoning

1 INTRODUCTION

Accurate trajectory prediction is paramount for the safety and efficiency of autonomous driving systems, and it demands a sophisticated understanding of complex traffic scenarios and human intentions (Geng *et al.*, 2023). The remarkable capabilities of large language models (LLMs) across various tasks, especially their potential in spatiotemporal forecasting and autonomous driving, spotlight the prospects of leveraging LLMs to enhance trajectory prediction endeavors. However, employing large language models in trajectory prediction necessitates addressing two critical issues: the integration of LLMs into this task and the mitigation of inference costs.

To overcome the above challenges, our study proposes the Chain-of-Thoughts Enhanced LLM-Instructed Trajectory Prediction Framework (LITPF), which utilizes an LLM to process textual traffic data within a Language-Instructed Module, thereby generating instrcutive texts. These texts are integrated with spatiotemporal data via a multimodal approach, enhancing the model's ability to interpret complex traffic dynamics and predict future movements more accurately. To tackle inference costs and improve LLM performance for trajectory prediction, we utilize existing trajectory prediction datasets with the GPT-4-Turbo LLM and a well-designed Chain-of-Thoughts (CoT) (Wei *et al.*, 2023) prompt engineering method to create textual traffic scenario datasets for fine-tuning a lightweight language model (LM). Additionally, we introduce a new spatial encoding technique and an enhanced uncertainty quantification method, to improve prediction accuracy and reliability. The workflow of the proposed framework is illustrated in Fig. 1 (b). The contributions of our research are significant and multifaceted:

(1) Establishment of the LITPF, a novel framework that integrates LLMs into trajectory prediction, enhancing model comprehension of dynamic traffic scenarios. This framework utilizes multimodal fusion of traffic data and instructive texts from LMs, introduces a novel spatial encoding method, and incorporates comprehensive uncertainty quantification to significantly enhance prediction precision and reliability.

(2) Development of the textualized traffic scenario datasets *Highway-Text* and *Urban-Text*, derived from real-world datasets using a CoT method and the GPT-4-Turbo. This significant contribution to the field facilitates the fine-tuning of lightweight LMs, optimizing computational efficiency and enhancing the practical utility of output texts for guiding trajectory predictions.

(3) Extensive experimental validation of the framework on multiple real-world datasets, demonstrating superior accuracy and reliability over existing methods, highlighting the framework's robustness and its potential to advance trajectory prediction technologies.

2 METHODOLOGY

2.1 Problem Formulation

Our research aims to predict the future trajectory of the *target vehicle* within a mixed autonomy environment, accounting for all traffic agents within the autonomous vehicle's (AV) sensing range. We focus on forecasting the trajectory $\boldsymbol{Y}$ of the target vehicle over a prediction horizon t_h based on its observed historical state $\boldsymbol{X}_{0:n}^{t-t_h:t}$ and those of surrounding traffic agents from $t - t_h$ to t.

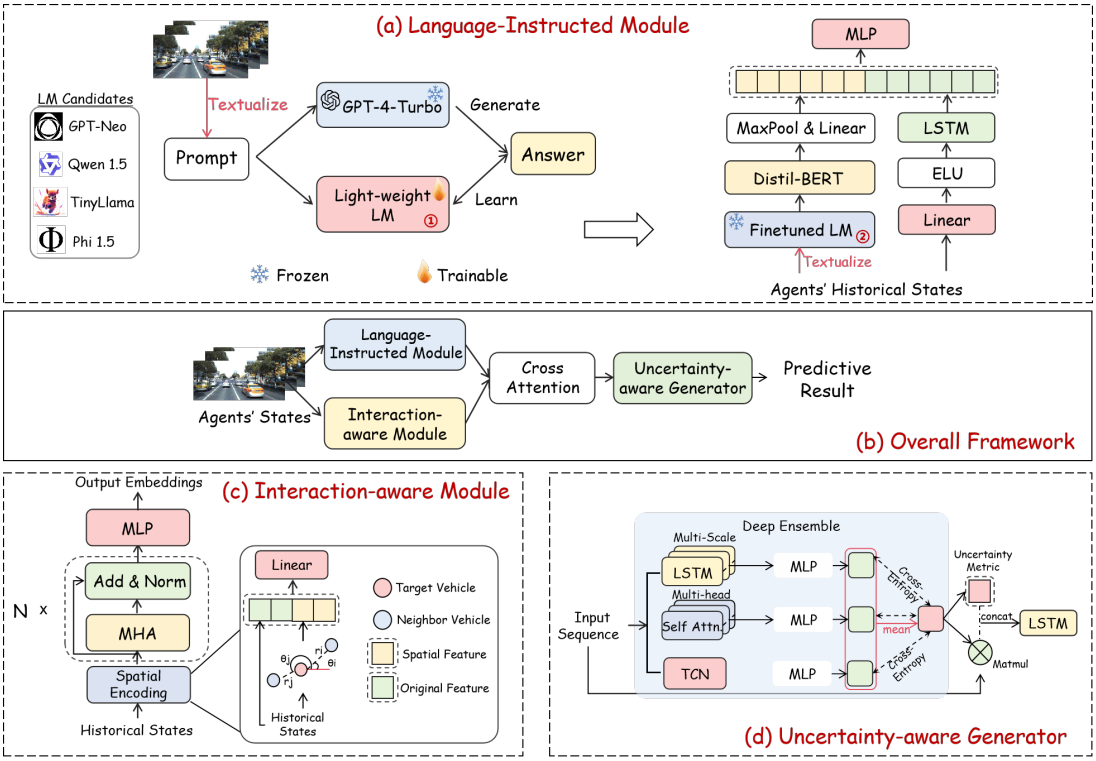

Figure 1 – *Illustration of the structure of the proposed LITPF: (a) details the workflow of the Language-Instructed Module; (b) depicts the overall framework architecture; (c) shows the structure of the Interaction-aware Module; and (d) illustrates the Uncertainty-aware Generator.*

2.2 Language-Instructed Module

Our methodology leverages LLMs as a source of heuristic information, guiding trajectory prediction through instructed textual insights. This capitalizes on the strengths of LLMs in generating contextually rich guidance, enhancing our model's ability to interpret complex traffic scenarios.

We leverage the capabilities of LLMs to perform zero-shot reasoning through carefully designed prompts. The dialogue with the LLM follows a logical progression, structured in a sequence that mirrors human cognitive processes: Interactions, Risks, and Predictions. In each thematic section, we enrich the query with commonsense knowledge and specific examples, aiding the LLM in its reasoning and response process. The LLM's response adheres to two key steps: initial reasoning followed by the provision of conclusive answers. Through our CoT approach, the LLM not only generates insightful observations but also stands out by detailing its reasoning process, thereby enhancing the interpretability and relevance of its responses. This structured

interaction paradigm is designed to optimize the LLM's performance in guiding trajectory prediction tasks, ensuring that it aligns with realistic and contextually accurate traffic scenarios. Details of our CoT-based method are available in Fig. 2.

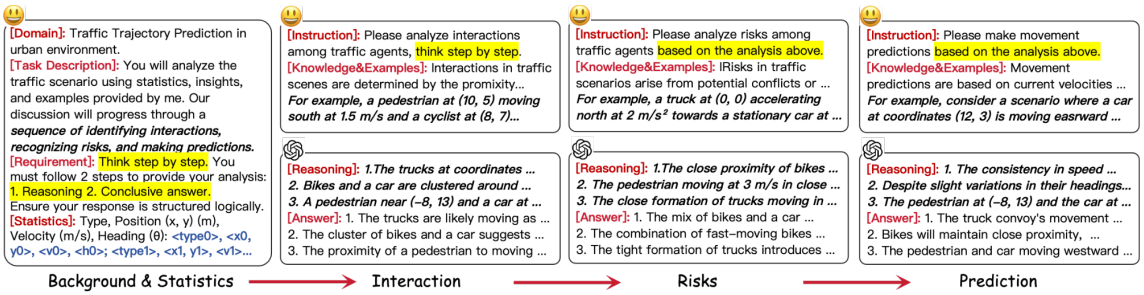

Figure 2 – *A detailed presentation of our CoT-based Prompt Engineering method.*

To mitigate the high operational costs and improve the practical application of LLMs in instructing trajectory prediction, we developed the *Highway-Text* and *Urban-Text* training datasets by sampling from constructed prompts derived from real-world traffic scenarios. Responses generated by the GPT-4-Turbo serve as labels for these datasets. Lightweight pretrained LMs are then fine-tuned on these datasets to generate the required instructional texts. To effectively utilize the generated instruction text, we employ the DistilBERT framework and max pooling to encode the textual information. This encoding is integrated with spatiotemporal information from the target agent through a multimodal fusion approach, as illustrated in Fig. 1 (a). This process ensures a comprehensive synthesis of language and spatiotemporal features, optimizing the prediction accuracy of our framework.

2.3 Interaction-aware Module

We've implemented the Transformer Encoder architecture to effectively capture complex interactions among agents, enhancing its understanding of spatial relationships through a new spatial encoding technique. At any given moment $t_k \in (t - t_h, t)$, with the target vehicle's position at $\boldsymbol{p}_0^{t_k}$ serving as the origin of the coordinate system, we compute the positions of other vehicles $\boldsymbol{p}_i^{t_k}, i \in (0, N]$ relative to the target vehicle in polar coordinates and get the displacement vector $\boldsymbol{d}_i^{t_k} = \boldsymbol{p}_i^{t_k} - \boldsymbol{p}_0^{t_k}$. Thus, the polar radius $r_i^{t_k}$, and the angle $\theta_i^{t_k}$ can be calculated. Subsequently, we concatenate the sets of angles $\Theta^{t_k} = \left\{\theta_i^{t_k} | i \in (0, N]\right\}$ and radial distances $\mathcal{R}^{t_k} = \left\{r_i^{t_k} | i \in (0, N]\right\}$ with the time frame's feature vector $\mathbf{X}^{t_k}$, followed by an embedding through a linear layer. Next, at each time step t_k, this feature is processed through a Transformer Encoder with shared weights, and subsequently projected using an MLP, as illustrated in Fig. 1 (c).

2.4 Uncertainty-aware Generator

In trajectory prediction for complex traffic scenarios, we address inherent uncertainties using a two-pronged approach: Aleatoric Uncertainty (AU) and Epistemic Uncertainty (EU). To address AU, we employ a Gaussian Mixture Model (Zhang & Li, 2022) to forecast 9 distinct maneuvers, which are categorized into 3 lateral and 3 longitudinal movements. This model calculates the maneuver state M and its probability based on observed states, which subsequently informs the trajectory prediction process. For EU, our strategy involves the use of a deep ensemble method incorporating Q distinct models, each offering unique data representations, as shown in Fig. 1 (d). This method diminishes model sensitivity to novel data by averaging the diverse probability distributions for the maneuver variable M across all models, thus enhancing the reliability of our predictions. We quantify EU using the average cross-entropy $\overline{H}(M)$ between the aggregated predictions and those from individual models, which helps refine the trajectory forecasts with bivariate Gaussian distribution parameters derived from these evaluations.

3 RESULTS

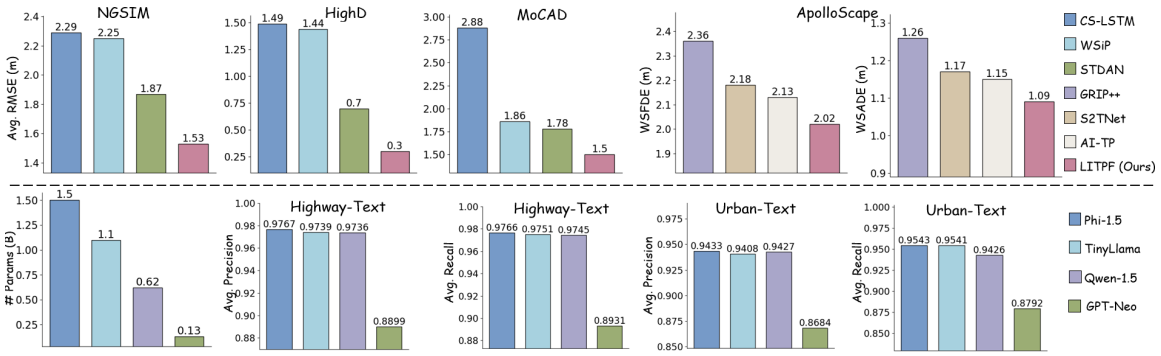

Figure 3 – *Experimental Results: The top column compares our model's performance against baseline models on the NGSIM, HighD, MoCAD, and ApolloScape datasets. The bottom one displays results for* 4 *LMs on Highway-Text and Urban-Text datasets, evaluated by BERT Score.*

In this study, the proposed LITPF underwent extensive evaluation across a range of real-world datasets, including NGSIM, HighD, MoCAD, and ApolloScape. The results, as depicted in Fig. 3 (top), demonstrate that LITPF achieves SOTA performance across several metrics in multiple real-world scenarios, validating the effectiveness of LITPF under varied conditions.

Additionally, we conducted tests on the specifically developed Highway-Text and Urban-Text datasets using four different sizes of LMs. The primary metric for evaluation was the BERT Score, with the results presented in Fig. 3 (bottom) for reference as baselines. The results indicate a significant improvement in performance from the 0.13B LM to the 0.62B LM. However, the difference in performance among the 0.62B, 1.1B and the 1.5B LMs is relatively minor. This suggests that smaller-sized models are capable of fulfilling the language-instructed tasks, indicating that increasing model size beyond a certain threshold yields diminishing returns.

4 DISCUSSION

Trajectory prediction is crucial for the safety of autonomous driving systems, and accurately predicting the movements of traffic agents is challenging. To tackle this, we introduced a framework LITPF, which leverages LLMs through CoT Reasoning method and multimodal fusion for instructing trajectory prediction. Additionally, this study contributes significantly to the field by developing the textualized traffic scenario datasets, enabling the fine-tuning of lightweight LMs generating instructive texts. The LITPF has demonstrated superior performance and robustness across 4 real-world datasets. Benchmark tests with differently parameterized LMs on our datasets confirm that lightweight LMs effectively handle this task, validating the feasibility and effectiveness of our approach.

References

Geng, Maosi, Chen, Yong, Xia, Yingji, & Chen, Xiqun (Michael). 2023. Dynamic-learning spatial-temporal Transformer network for vehicular trajectory prediction at urban intersections. *Transportation Research Part C: Emerging Technologies*, **156**, 104330.

Wei, Jason, Wang, Xuezhi, Schuurmans, Dale, Bosma, Maarten, Ichter, Brian, Xia, Fei, Chi, Ed, Le, Quoc, & Zhou, Denny. 2023. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.*

Zhang, Kunpeng, & Li, Li. 2022. Explainable multimodal trajectory prediction using attention models. *Transportation Research Part C: Emerging Technologies*, **143**, 103829.