2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

ESP: Extro-Spective Prediction for Long-term Behavior Reasoning in Emergency Scenarios

Dingrui Wang[1,4], Zheyuan Lai[1], Yuda Li[1], Yi Wu[2], Yuexin Ma[3], Johannes Betz[4],
Ruigang Yang[1], *Fellow, IEEE*, Wei Li[1]

***Abstract*— Emergent-scene safety is the key milestone for fully autonomous driving, and reliable on-time prediction is essential to maintain safety in emergency scenarios. However, these emergency scenarios are long-tailed and hard to collect, which restricts the system from getting reliable predictions. In this paper, we build a new dataset, which aims at the long-term prediction with the inconspicuous state variation in history for the emergency event, named the Extro-Spective Prediction (ESP) problem. Based on the proposed dataset, a flexible feature encoder for ESP is introduced to various prediction methods as a seamless plug-in, and its consistent performance improvement underscores its efficacy. Furthermore, a new metric named clamped temporal error (CTE) is proposed to give a more comprehensive evaluation of prediction performance, especially in time-sensitive emergency events of subseconds. Interestingly, as our ESP features can be described in human-readable language naturally, the application of integrating into ChatGPT also shows huge potential. The ESP-dataset and all benchmarks are released at `https://dingrui-wang.github.io/ESP-Dataset/`.**

I. INTRODUCTION

Autonomous driving (AD) is attracting significant attention as well as a large amount of investment and resources. However, the business success of AD is still stuck at level 2/3. The driverless solution, *i.e.* the level-4 AD, has slowed its pace down towards mass production as safety must be verified with zero-tolerance solidly. To improve the safety of the AD system, one of the most key technologies is prediction. An acute yet reliable algorithm to predict the future states of the surrounding traffic participants is the cornerstone of safe decision and driving control.

The rapid advance of prediction, especially powered by artificial intelligence techniques, has been witnessed in recent years. The seed work of the deep learning-based prediction algorithm is LSTM [1], [15]. It encodes the states, *e.g.* the velocity and position of traffic participants, in the past few seconds to estimate the future states. We name such historical states based method to *introspective prediction*, which heavily relies on information in its own right. However, such introspective prediction methods suffer from multi-agent interaction. The pioneer work taking the influence of other traffic participants into account is social LSTM [16]. An inspiring work in this line of research is VectorNet [13], which encodes the interaction between not only the ego with the agents but also the ego with HDMap, the static environment information. Such lightweight HDMap vectorization technique brings huge performance improvement w.r.t. the displacement error of predicted trajectories.

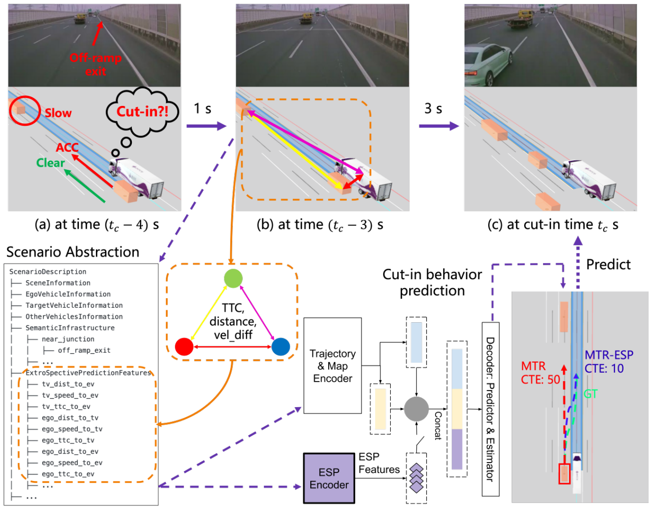


Fig. 1: A real emergency scenario with a sedan dangerously cut-in in front of the AD truck on the highway. At the time in (a), human drivers foretell sedan's behavior by interpreting *extrospective* cues: 1) *[observe]* a high-speed accelerating (**ACC**) sedan approaching a **Slow** front-blocking truck, *[predict]* high potential left/right lane change and low possibility of hard brake for the sedan, 2) *[observe]* left lane of the sedan is **Clear**, *[predict]* left lane change will not happen as it can be done at anytime earlier with lower risk. 3) *[observe]* an **Off-ramp exit** in about 200 meters, *[predict]* likely to force cut-in into far-right lane to catch exit. Note the sedan did exit the highway as expected in this case. The MTR method predicts the behavior at (c). While the ESP encoder can absorb the *extrospective* cues to predict the cut-in event in advance as shown in (b).

Even with the social interaction encoding, the SOTA prediction methods [35], [43] are still not as intelligent as a human and always fail when facing complex or emergency scenarios, which requires a deep understanding of the environment and multiple lines of reasoning. Fig. 1 shows a real-world *emergency cut-in* case from our AD fleet. The

---

*The study was supported by the Natural Science Foundation of Jiangsu Province (BK20210600)

[1] Inceptio Technology, Shanghai 200082, China, `dingrui.wang, zheyuan.lai, yuda.li, yang.ruigang, wei.li @inceptio.ai`

[2] Nanjing University of Posts and Telecommunications, Nanjing 210023, China, `yiw@njupt.edu.cn`

[3] ShanghaiTech University, Shanghai 200120, China, `mayuexin@shanghaitech.edu.cn`

[4] The authors are with the Professorship of Autonomous Vehicle Systems, Technical University of Munich, 85748 Garching, Germany; Munich Institute of Robotics and Machine Intelligence (MIRMI), `dingrui.wang@hotmail.com, johannes.betz@tum.de`

979-8-3503-8457-4/24/$31.00 ©2024 IEEE

TABLE I: A comprehensive comparison of motion prediction datasets.

Dataset	Year	Segments	Time horizon	Sampling rate	Boxes	Distance	Density of aggressive behavior	Semantic map	Highway	Offline perception	Semantic environment information
Lyft [22]	2019	170k	5s	10 Hz	2D	10 km	low				
TrafficPredict [27]	2019	-	-	10 Hz	-	-	high				
NuScenes [5]	2020	1k	6s	2 Hz	3D	-	low				
Argoverse [6]	2019	324k	3s	10 Hz	None	290 km	low	✓			
INTERACTIONS [42]	2019	-	3s	10 Hz	2D	-	high	✓	✓	✓	
WOMD [11]	2021	104k	8s	10 Hz	3D	1750 km	low	✓	✓	✓	
ESP-Dataset	2023	120k	5s	10 Hz	2D	2100 km	very high	✓	✓	✓	✓

human driver can predict the emergency cut-in event at the most early timing shown in Fig. 1 (a). The advanced MTR method [35] only yields the prediction when lateral movement is observable at the timing shown in Fig. 1 (c).

The lesson learned from this real case is the huge gap between human extrospective cue understanding and reasoning ability and the performance of current prediction algorithms. Rethinking the MTR/TNT and related works [13], [26], [34], [39], even attention mechanism incorporating with local/global graph is well-designed to encode historical state and HDMap information, the global attention on extrospective cue and corresponding inference ability is really weak. Among the reasons is the deficiency of the dataset focusing on rare and difficult-to-predict scenarios with rich but often-overlooked environmental information. Furthermore, a more proper metric than Final Displacement Error (FDE) and Average Displacement Error (ADE), as well as powerful baselines should be developed for those challenge scenarios.

In this paper, we tackle the problem of long-term prediction in emergency scenarios, where internal states in history are inconspicuous but the external environment provides effective cues for reasoning in the best time window ahead of the emergency event. We name it as the extrospective problem (ESP). Due to the lack of related datasets, we built the ESP-Dataset for challenging scenarios with emergency events. Moreover, the dataset is collected and labeled for diverse scenarios over 2k+ kilometers, with novel and unique semantic environment information for extrospective prediction provided. Furthermore, a new metric named Clamped Temporal Error(CTE) is designed to comprehensively evaluate time-wise prediction performance, which is an important but missing aspect for sub-second-level time-sensitive emergency scenarios. Lastly, ESP feature extraction and network encoder are introduced, which would benefit existing backbones and algorithms seamlessly. Back to the emergency case in Fig. 1, by introducing ESP encoding, MTR with ESP effectively forecasts the cut-in three seconds in advance than MTR shown in Fig. 1 (b). One more interesting thing is that our ESP features can be described in human-readable language naturally, and we have already integrated it into large language models such as GPT [29] .

The contributions of the paper are summarized below:

- The ESP-Dataset with semantic environment information is collected over 2k+ kilometers focusing on emergency-event-based challenging scenarios.
- A new metric named CTE is proposed for comprehensive evaluation of prediction performance in time-sensitive emergency scenarios.
- ESP feature extraction and network encoder are introduced, which can be used to enhance existing backbones/algorithms seamlessly.

II. RELATED WORK

A. Motion prediction datasets

In the last decade, significant advancements in Autonomous Vehicle (AV) have been propelled by the availability of diverse datasets for understanding scenes. This journey began with 2D annotated datasets (such as CamVid [4] and Apolloscape [17]), which evolved into multimodal datasets (like KITTI [14] and KAIST [8]), incorporating images, range sensor data (lidars, radars), and GPS/IMU data *etc*.. However, these datasets primarily focus on *introspective* cues, as they only consider historical information, neglecting semantic details. In addition, numerous other motion prediction datasets also have been established including the Stanford Drone Dataset [33], Town Center [3], NGSIM [10], ETH [30], Automatum [36], UCY [24], highD [23], and exiD [28]. It is worth noting that while these datasets offer valuable insights into motion prediction, they primarily focus on specific fixed locations rather than the broader context of dynamic driving environments. Moving on to large-scale AV datasets, nuScenes [5], Lyft L5 [22], TrafficPredict [27], Argoverse [6], INTERACTION [42], and Waymo Open Motion dataset (WOMD) [11] have been made available to the public. The INTERACTION dataset selects particular driving locations (e.g., roundabouts) to emphasize interactive complexity, while WOMD aims to jointly predict motion behavior. As previously mentioned, understanding context requires considering semantic information related to *extrospective* aspect. TrafficPredict, Argoverse, INTERACTION, and WOMD provide HD maps rich in semantics. Nevertheless, these datasets can only provide *introspective* cues to prediction models, neglecting *extrospective* factors. NuScenes and Lyft L5 dataset differ in that they do not focus on interactive driving scenarios [5]. An overall comparison of various AV datasets is presented in Table I.

B. Driving behavior prediction

Theory of Mind (ToM), the ability to understand the mental states of others, is a primary reason humans can successfully negotiate traffic on a highway onramp [18]. As previously mentioned, existing models such as Vectornet [13], TNT [43], Grip++ [26], Graph-based network [34], GANet [39] and MTR [35], *etc*. tend to focus on *introspective* aspect, which relies heavily on historical data and high-definition maps. Furthermore, their performance is still far from human-level interpretation of semantic environment information, and struggle with complex scenarios that require a deep understanding of the environment through complex reasoning [20]. This is precisely why the advancement of current models requires richer and more detailed semantic information. In light of this, we introduce the ESP encoder to demonstrate the seamless integration capabilities of the ESP dataset. Large Language Models (LLMs) such as the chatGPT [29] and GPT-4 [31] have been increasingly gaining attention recently. And there are recent studies have explored their applications in autonomous driving tasks [12]. Our dataset is also compatible with LLMs regarding input format.

Concerning the prediction of cut-in behaviors, various attempts have been made such as convolution neural network-based method [21] and behavioral probability distribution-based method [19]. However, existing metrics to evaluate the prediction result such as Time-to-Collision [28], Final Displacement Error (FDE), Average Displacement Error (ADE), MR (Missing Rate) [11] and Average Precision [5] do not adequately assess the quality of cut-in predictions. Specifically, traditional metrics evaluate cut-in predictions based on trajectory rather than the precise cut-in moment. Consequently, current metrics emphasize spatial information while neglecting temporal aspects during evaluation.

Fig. 2: Sensor setup for the ESP data collection platform involves the Inceptio autonomous truck, which is equipped with 5 LiDARs, 7 cameras, 7 radars, and GPS.

III. THE ESP-DATASET

A. Sensor Setup

We use Inceptio autonomous trucks [37] to capture data as shown in Fig. 2. The sensor system of the Inceptio autonomous truck includes 5 LiDARs, 7 cameras, 7 radars, and GPS. All five LiDARs are solid-state. One forward long-range LiDAR is above the front windshield, two rear-facing long-range LiDARs are under the rearview mirrors on both sides, and two short-range blind zone LiDARs are on both sides of the roof. Among them, long-range LiDARs can detect up to 200 meters with 120 degrees FOV and short-range LiDARs can scan up to 140 degrees. Three cameras with short, middle, and long ranges are mounted on the top of the windshield. The other four cameras are installed above the truck doors on both sides, and these cameras are forward middle-range cameras and backward fish-eye cameras. Five long-range radars are placed on the bumper with forward, left-forward, right-forward, left-rear, and right-rear views. The two other two long-range radars are mounted under the rearview mirrors. GPS is installed on the roof.

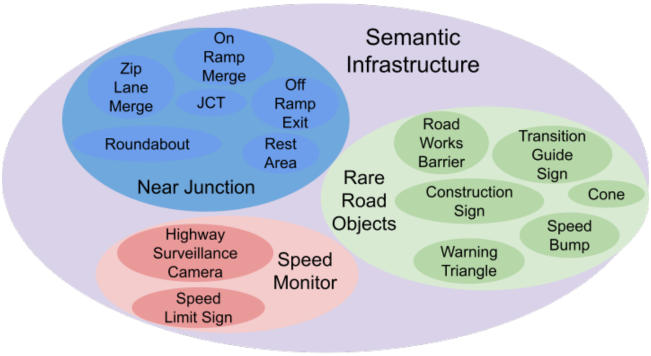


Fig. 3: Organization of Semantic Infrastructure. The figure illustrates the organization of Semantic Infrastructure, which comprises three extrospective components: speed monitoring systems, junctions, and rare road objects.

B. Scenario Description Paradigm

A complete representation of the agent's operating environment is critical for testing and evaluating autonomous driving models [25], [32]. Different models use various representations to describe the driving scenario. In MDP and MCTS [2], [40], it's a state space with vehicle position, velocity, and actions. Lev (Lane-based planning) [38] uses lane information while Free-space velocity (FV) [9] has no strict lane constraints. Social LSTM [16] uses the trajectories of multiple agents. For Graph-based methods [26], [34], agents are nodes, and interactions are edges. Moreover, there are models [35], [39], [43] that focus on using agent trajectories and map polylines as the encoder's input. However, previous models lack a semantics-oriented *extrospective* understanding of scenario descriptions. Large Language Models (LLMs) [12], [41] are capable of absorbing natural language-oriented information and providing comprehensive insights. Similarly, a driving scenario can be organized in a format that incorporates natural language information.

Inspired by this, to describe the scenario in a complete manner, the paradigm of the scenario description to build the dataset includes the following components.

- **Scene** covers diverse attributes like lane type, weather conditions, total vehicles in scope, etc.

- **Ego vehicle**'s lane location together with its historical velocity and trajectory data are given.
- **Target vehicle** is the object for which future behavior needs to be predicted. This section includes information presented in a format similar to that of the ego vehicle, with the distinction that it also includes the ground truth of its future behavior.
- **Relative Interaction Vehicles** describes surrounding vehicles in a format similar to that of the ego vehicle.
- **Semantic Infrastructure** encompasses three *extrospective* components: speed monitoring systems, junctions and rare road objects. The detailed framework of the semantic infrastructure is illustrated in Fig.3.
- ***Extrospective* Features** contains different features related to the distance and relative longitudinal velocity between different agents in an ESP token.

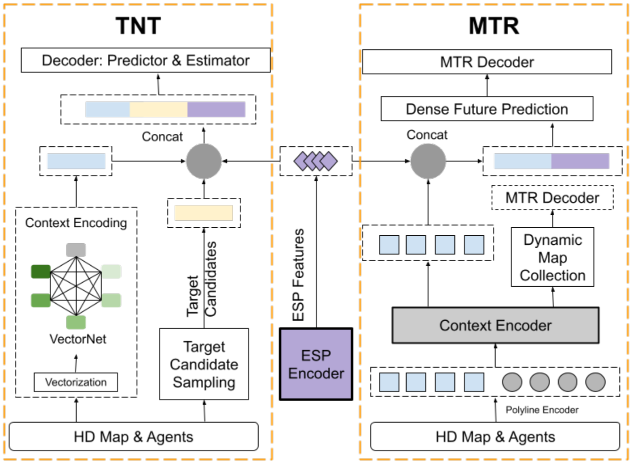


Fig. 4: Seamless Integration of ESP Features with Motion Prediction Models. The ESP plugin seamlessly integrates with widely used encoder-decoder models such as TNT and MTR. As depicted, ESP enhances existing features through straightforward concatenation, leading to a transformative advancement in motion prediction.

C. Easy plug-in ESP Encoder

We propose to extract ESP features with a simple network encoder. Even though the encoder in our case is as simple as a one-layer standardized LSTM module, the improvement to the performance of the SOTA models is already considerable. The related experiment results will be shown in section V. And the results have shown the easy plug-in property of the ESP encoder which can also enhance existing backbones/algorithms seamlessly. As depicted in Fig. 4, the result generated by the ESP encoder is directly concatenated to the original encoder output of the model. The only aspect to be mindful of is to account for the input dimension of the model's decoder.

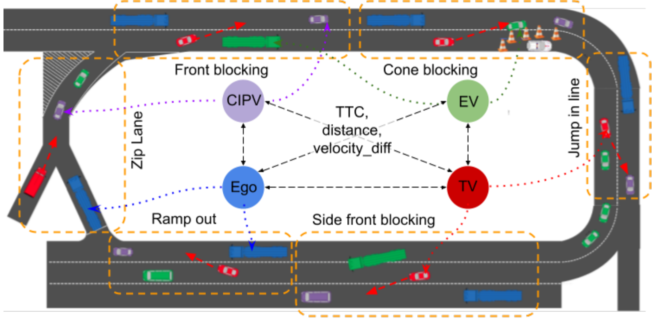


Fig. 5: ESP Token Types - Representation of Ego (Ego vehicle), CIPV (Vehicle in front of ego), EV (Environmental vehicle), and TV (Target vehicle) within each time frame. Scenario types are determined based on rule-based criteria.

D. Data mining

As depicted in Fig. 5, collected scenarios encompassed various interactions such as merges, lane changes, ramp out, cone block, and zip lane. Interesting scenarios were mined token by token and based on different spatial-temporal criteria which are performed every three frames, using the current frame at time t_0, the historical frame at time t_{-3}, and the future frame at time t_3 as a base for detection.

An example of the mining detection criteria for front-blocking scenarios is: there is a slow-moving vehicle agent A ahead, and the Time-to-Collision is less than 5 seconds. In the next 5 seconds, if the ego vehicle's minimum deceleration is heavier than a threshold (*e.g.*-0.9 m/s^2) or the average deceleration is heavier than a threshold (*e.g.*-0.5 m/s^2), and the preceding vehicle either performs a cut-in maneuver or is close to the ego vehicle's dangerous zone.

Once these conditions are satisfied, a token for the abstraction of a scenario will be extracted. The mined token result is illustrated in Fig. 6, the token spans over a history of three seconds and includes a ground truth trajectory for the target vehicle to be predicted over five seconds. Therefore, the overall time horizon is eight seconds.

The map data covers the Shanghai-Jinan and Shanghai-Quanzhou highways, as well as the Shanghai Outer Ring Road. We have a representation of lane features including the lane centerlines, lane boundary lines, and road edges.

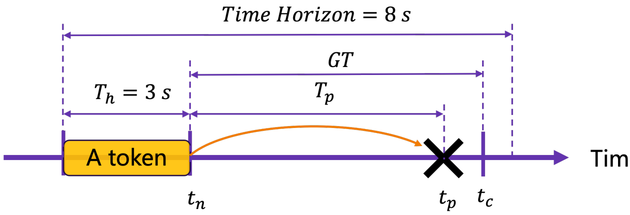


Fig. 6: This figure depicts the framework of a token. The token spans a history of three seconds and includes a ground truth trajectory for the target vehicle to be predicted over five seconds. Therefore, the overall time horizon is eight seconds. Here, t_n represents the current time stamp, t_p is the predicted time stamp, and t_c denotes the ground truth cut-in moment.

E. Statistics

ESP covers more than 110,000 tokens on highways with a total length of around 2100 km. The annotation frequency is 10Hz. We organize the data by frames, with each frame containing all the traffic agents' IDs, categories, positions, and bounding boxes. We divide the dataset into training, validation, and testing sets by 8:1:1. Fig. 7 provides 8 typical scenarios sampled in the ESP-dataset. It can be seen that the traffic in our dataset features on challenging scenario that has high density, the surrounding areas of the ego-truck are occupied by heterogeneous vehicles, such as cars and trucks.

(a) Side front blocking (b) Zip lane (c) Ramp out (d) Jump in line

(e) Speed limit camera (f) Cone blocking (g) front blocking (h) Transition guide

Fig. 7: Example scenarios from the ESP-dataset captured by the front and side cameras of our ego-truck.

IV. TASK AND METRIC

A. *Cut-in Evaluation Dilemma*

Regarding driving behavior, with a specific focus on cut-in behavior, existing models primarily centered on trajectory prediction can also provide predictions for cut-in behavior. This can be accomplished by superimposing the bounding box onto the predicted trajectory while considering the heading for each point along the trajectory, the initial intersection point between the bounding box and the lane's polyline can be identified. Despite the straightforward conversion of trajectory prediction into behavior prediction, assessing the prediction outcomes based on current metrics such as Final Displacement Error (FDE), Average Displacement Error (ADE) and MR (Missing Rate) poses a non-trivial challenge. The reason behind the challenge is that these metrics focus on overall trajectory performance without considering temporal details. For example, in the Fig. 8, trajectories labeled "a" and "b" would yield the same results with these metrics regarding the ground truth in between, despite "a" showing an earlier lane change. These observations highlight the limitations of current metrics and emphasize the need to capture the temporal aspects of prediction behaviors for a more comprehensive evaluation.

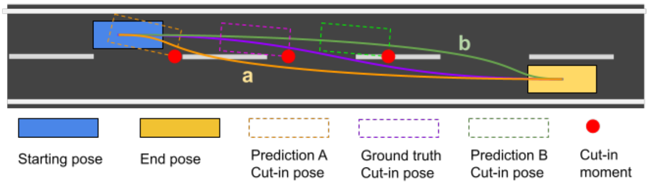


Fig. 8: Illustration of the necessity to consider CTE metric in the assessment of cut-in behavior prediction.

B. *Metric*

We introduce the Clamped Temporal Error (***CTE***) to measure the difference between predicted and actual behavior times. For each scenario token S that we evaluate, our model generates K potential predictions, denoted as P_k, $k \in 1 \ldots K$. Each prediction P_k is related to a trajectory $s_k = \{s_{a,t}\}_{t=1:T, a=1:A}$ for T future time steps for A agents. Similarly, the ground truth is denoted as $\hat{s} = \{\hat{s}_{a,t}\}$. The individual object prediction task becomes a special case of this formulation where each joint prediction contains only a single agent $A = 1$. The ***minCTE*** is computed as formulated in the equation below,

$$\boldsymbol{minCTE} = \min_k \sum_{a=1}^{A} \min(||LaMT(\hat{s}_{a,t}) - LaMT(s^k_{a,t})||, t_u) \tag{1}$$

where $LaMT$ represents the function for calculating Lane Match Time by determining the intersection point between the input trajectory and the polylines of nearby lanes w.r.t. the vehicle's heading and boundary box. The term t_u refers to the upper threshold to clamp the time difference.

TABLE II: Ablation study on ESP features using TNT, retrained TNT, and MTR with ESP encoder.

Model	minFDE	minADE	minCTE	Precis.	Recall	Acc.
TNT-ESP	***4.33***	***2.00***	***0.71***	***0.75***	***0.83***	***0.76***
w/o tv-ev	*4.89*	*2.21*	*0.74*	*0.73*	*0.83*	*0.75*
w/o tv-cipv	*5.22*	*2.32*	*0.75*	*0.75*	*0.80*	*0.75*
w/o ego-tv	*5.88*	*2.59*	*0.82*	*0.72*	*0.78*	*0.72*
w/o ego-cipv	*4.87*	*2.20*	*0.74*	*0.73*	*0.83*	*0.74*
w/o ego-ev	*5.23*	*2.33*	*0.75*	*0.72*	*0.82*	*0.74*
TNT-base	*5.21*	*2.33*	*0.75*	*0.74*	*0.81*	*0.74*
TNT-ESP (retrained)						
w/o tv-ev	*4.46*	*2.06*	*0.75*	*0.72*	*0.84*	*0.74*
w/o tv-cipv	*4.45*	*2.06*	*0.73*	*0.73*	*0.85*	*0.75*
w/o ego-tv	*4.46*	*2.05*	*0.73*	*0.73*	*0.84*	*0.75*
w/o ego-cipv	*4.52*	*2.08*	*0.73*	*0.73*	*0.84*	*0.75*
w/o ego-ev	*4.44*	*2.04*	*0.75*	*0.72*	*0.85*	*0.74*
TNT-base	*4.52*	*2.07*	*0.74*	*0.74*	*0.83*	*0.75*
MTR-ESP	***1.04***	***0.56***	***0.23***	***0.72***	***0.955***	***0.932***
MTR-base	*1.06*	*0.58*	*0.24*	*0.70*	*0.953*	*0.930*

V. ESP APPLICATIONS

A. *The Effectiveness of ESP Encoder*

The ESP encoding has demonstrated its significant value by producing rapid and discernible improvements in model performance utilizing the ESP features from the ESP-dataset. As shown in Table II, our initial experiment with TNT [43] serves to elucidate the distinct influences exerted by each component of the ESP encoder. It is evident that the ESP encoder substantially enhances performance in nearly all aspects. Furthermore, our ablation study on TNT, retrained with only different partial ESP features, reveals that different segments of the ESP features contribute to varying degrees of influence on the outcome. Additionally, in our experiment involving MTR [35], the results demonstrate that the ESP encoder can still enhance the state-of-the-art (SOTA) baseline's performance considerably. In Fig. 9, three cases are presented to demonstrate the substantial improvement in the model's performance for cut-in behavior prediction achieved through MTR with the use of the ESP encoder. As in Case 1.1, the ground truth cut-in moment occurs at 2.2 seconds, whereas the MTR-base model predicts it at 4.4 seconds (Case

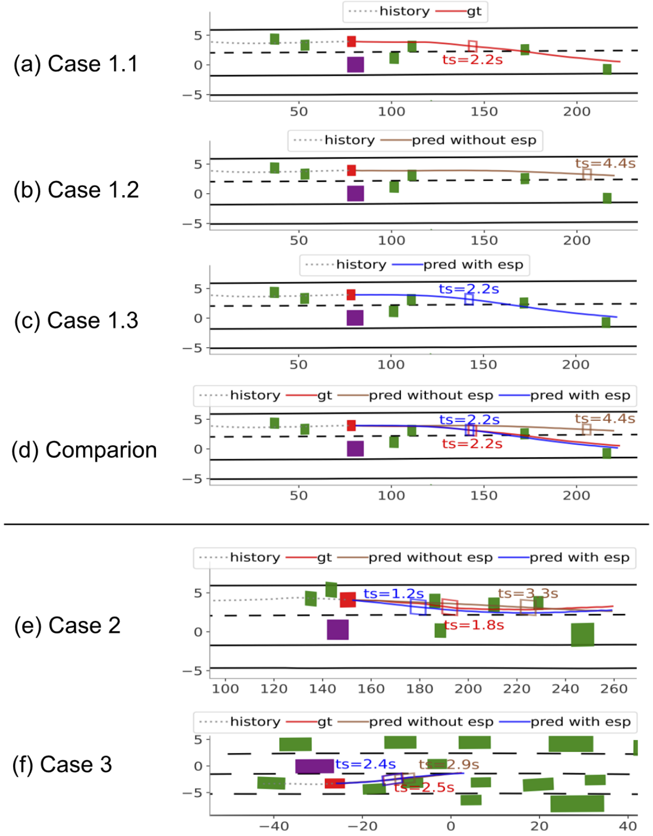
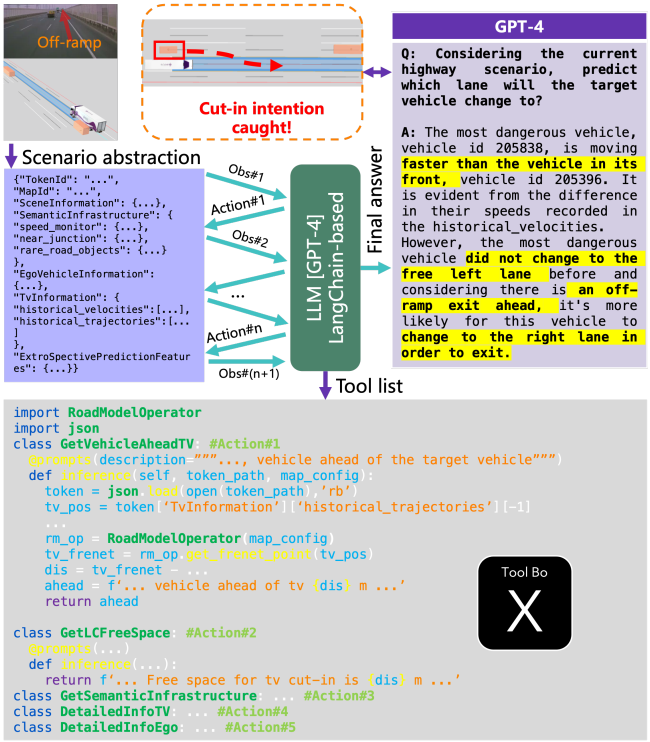


Fig. 9: MTR-base and MTR-ESP prediction performance comparison. Case 1.1 to Case 1.4 are the ground truth trajectory for the cut-in behavior, the prediction results of the MTR-base model without ESP features, the prediction results of the MTR model with the ESP encoder, and a combined plot respectively.

1.2). On the contrary, the MTR-ESP model provides a more accurate prediction of 2.2 seconds (Case 1.3).

B. *Friendly to LLM*

The GPT model [31] has gained widespread attention in recent months. The emergence of ChatGPT [29] has been captivating the world. Naturally, this leads to the question of how GPT can contribute to the field of motion prediction. In this paper, we explore one potential application of the ESP-dataset: utilizing its tokens directly as input for a Large Language Model. We specifically evaluate its performance within the context of the widely endorsed LangChain framework [7], as recent studies have advocated for its application in prediction tasks [12]. The model pipeline commences by converting the highway scenario into a standardized format and feeding it into GPT-4 through the LangChain interface. Subsequently, the LLM provides insights through multiple action-observation pairs. Actions are defined within the "toolbox". Once GPT believes that it has gathered enough observational information, it provides the final answer. The driving scenario mentioned in the introduction section goes through this LangChain-based pipeline, with details presented in Fig.10.



```
{"TokenId": "...",
"MapId": "...",
"SceneInformation": {...},
"SemanticInfrastructure": {
"speed_monitor": {...},
"near_junction": {...},
"rare_road_objects": {...}
},
"EgoVehicleInformation":
{...},
"TvInformation": {
"historical_velocities":[...],
"historical_trajectories":[...
]
},
"ExtroSpectivePredictionFeatur
es": {...}}
```

Q: Considering the current highway scenario, predict which lane will the target vehicle change to?

A: The most dangerous vehicle, vehicle id 205838, is moving **faster than the vehicle in its front,** vehicle id 205396. It is evident from the difference in their speeds recorded in the historical_velocities. However, the most dangerous vehicle **did not change to the free left lane** before and considering there is **an off-ramp exit ahead,** it's more likely for this vehicle to **change to the right lane in order to exit.**

```
import RoadModelOperator
import json
class GetVehicleAheadTV: #Action#1
  @prompts(description="""..., vehicle ahead of the target vehicle""")
  def inference(self, token_path, map_config):
    token = json.load(open(token_path),'rb')
    tv_pos = token['TvInformation']['historical_trajectories'][-1]
    ...
    rm_op = RoadModelOperator(map_config)
    tv_frenet = rm_op.get_frenet_point(tv_pos)
    dis = tv_frenet - ...
    ahead = f'... vehicle ahead of tv {dis} m ...'
    return ahead

class GetLCFreeSpace: #Action#2
  @prompts(...)
  def inference(...):
    return f'... Free space for tv cut-in is {dis} m ...'
class GetSemanticInfrastructure: ... #Action#3
class DetailedInfoTV: ... #Action#4
class DetailedInfoEgo: ... #Action#5
```

Fig. 10: Demonstration of highway scenario processing using the LangChain-based pipeline with GPT-4. GPT-4 initiates with the 'GetVehicleAheadTV' tool and receives an observation indicating the presence of a small vehicle ahead of the target vehicle (TV). Following further inferences, the model ultimately predicts the cut-in intention of the target vehicle in advance. Further, the explanations provided by the model are similar to human thought processes.

VI. CONCLUSION AND FUTURE WORK

In this paper, we addressed the critical challenge of long-term prediction in autonomous driving, with a focus on emergency cut-in scenarios where semantic environmental cues are pivotal. We introduced the ESP problem, curated the ESP-Dataset enriched with semantic environment information, and introduced the Clamped Temporal Error (CTE) metric for time-sensitive emergency scenario assessment. Our ESP feature extraction with the ESP encoder significantly boosted existing prediction methods, particularly in complex interaction scenarios, as evidenced by TNT and MTR model experiments. Furthermore, we unveiled the potential of incorporating ESP features into large language models like GPT to make better predictions by using *extrospective* cues.

In the future, we plan to make the ESP-dataset available to the research community and also continue expanding the ESP-Dataset by including more diverse and challenging scenarios. Additionally, our research will focus on developing advanced ESP encoders that incorporate causal reasoning techniques for further improvement. Furthermore, fine-tuning LLMs with ESP features, aiming to improve scenario understanding for autonomous vehicles, is an interesting direction.

REFERENCES

[1] Florent Altché and Arnaud de La Fortelle. An lstm network for highway trajectory prediction. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 353–359. IEEE, 2017.

[2] Haoyu Bai, Shaojun Cai, Nan Ye, David Hsu, and Wee Sun Lee. Intention-aware online pomdp planning for autonomous driving in a crowd. In *2015 ieee international conference on robotics and automation (icra)*, pages 454–460. IEEE, 2015.

[3] Ben Benfold and Ian Reid. Stable multi-target tracking in real-time surveillance video. In *CVPR 2011*, pages 3457–3464. IEEE, 2011.

[4] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10*, pages 44–57. Springer, 2008.

[5] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[6] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8748–8757, 2019.

[7] Harrison Chase. Langchain. *Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, HP d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph*, 2022.

[8] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018.

[9] Laurene Claussmann, Marc Revilloud, Dominique Gruyer, and Sébastien Glaser. A review of motion planning for highway autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 21(5):1826–1848, 2019.

[10] Benjamin Coifman and Lizhe Li. A critical evaluation of the next generation simulation (ngsim) vehicle trajectory dataset. *Transportation Research Part B: Methodological*, 105:362–377, 2017.

[11] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021.

[12] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive like a human: Rethinking autonomous driving with large language models. *arXiv preprint arXiv:2307.07162*, 2023.

[13] Jiyang Gao, Chen Sun, Hang Zhao, Yi Shen, Dragomir Anguelov, Congcong Li, and Cordelia Schmid. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11525–11533, 2020.

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[16] Lian Hou, Long Xin, Shengbo Eben Li, Bo Cheng, and Wenjun Wang. Interactive trajectory prediction of surrounding road users for autonomous driving using structural-lstm network. *IEEE Transactions on Intelligent Transportation Systems*, 21(11):4615–4625, 2019.

[17] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2702–2719, 2019.

[18] Boris Ivanovic and Marco Pavone. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2375–2384, 2019.

[19] Rubén Izquierdo, A Quintanar, Ignacio Parra, D Fernández-Llorca, and MA Sotelo. The prevention dataset: a novel benchmark for prediction of vehicles intentions. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 3114–3121. IEEE, 2019.

[20] Phillip Karle, Maximilian Geisslinger, Johannes Betz, and Markus Lienkamp. Scenario understanding and motion prediction for autonomous vehicles—review and comparison. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):16962–16982, 2022.

[21] Hadi Kazemi, Hossein Nourkhiz Mahjoub, Amin Tahmasbi-Sarvestani, and Yaser P Fallah. A learning-based stochastic mpc design for cooperative adaptive cruise control to handle interfering vehicles. *IEEE Transactions on Intelligent Vehicles*, 3(3):266–275, 2018.

[22] R Kesten, M Usman, J Houston, T Pandya, K Nadhamuni, A Ferreira, M Yuan, B Low, A Jain, P Ondruska, et al. Lyft level 5 perception dataset 2020, 2019.

[23] Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st international conference on intelligent transportation systems (ITSC)*, pages 2118–2125. IEEE, 2018.

[24] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007.

[25] Wei Li, CW Pan, Rong Zhang, JP Ren, YX Ma, Jin Fang, FL Yan, QC Geng, XY Huang, HJ Gong, et al. Aads: Augmented autonomous driving simulation using data-driven algorithms. *Science robotics*, 4(28):eaaw0863, 2019.

[26] Xin Li, Xiaowen Ying, and Mooi Choo Chuah. Grip++: Enhanced graph-based interaction-aware trajectory prediction for autonomous driving. *arXiv preprint arXiv:1907.07792*, 2019.

[27] Yuexin Ma, Xinge Zhu, Sibo Zhang, Ruigang Yang, Wenping Wang, and Dinesh Manocha. Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6120–6127, 2019.

[28] Tobias Moers, Lennart Vater, Robert Krajewski, Julian Bock, Adrian Zlocki, and Lutz Eckstein. The exid dataset: A real-world trajectory dataset of highly interactive highway scenarios in germany. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 958–964. IEEE, 2022.

[29] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[30] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th international conference on computer vision*, pages 261–268. IEEE, 2009.

[31] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[32] Rodrigo Queiroz, Thorsten Berger, and Krzysztof Czarnecki. Geoscenario: An open dsl for autonomous driving scenario representation. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 287–294. IEEE, 2019.

[33] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 549–565. Springer, 2016.

[34] Zihao Sheng, Yunwen Xu, Shibei Xue, and Dewei Li. Graph-based spatial-temporal convolutional network for vehicle trajectory prediction in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):17654–17665, 2022.

[35] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. *Advances in Neural Information Processing Systems*, 35:6531–6543, 2022.

[36] Paul Spannaus, Peter Zechel, and Kilian Lenz. Automatum data: Drone-based highway dataset for the development and validation of automated driving software for research and commercial applications. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 1372–1377. IEEE, 2021.

[37] Dawei Wang, Lingping Gao, Ziquan Lan, Wei Li, Jiaping Ren, Jiahui Zhang, Peng Zhang, Pei Zhou, Shengao Wang, Jia Pan, et al. An intelligent self-driving truck system for highway transportation. *Frontiers in neurorobotics*, 16:843026, 2022.

[38] Jingke Wang, Tengju Ye, Ziqing Gu, and Junbo Chen. Ltp: Lane-based trajectory prediction for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17134–17142, 2022.

[39] Mingkun Wang, Xinge Zhu, Changqian Yu, Wei Li, Yuexin Ma, Ruochun Jin, Xiaoguang Ren, Dongchun Ren, Mingxu Wang, and

Wenjing Yang. Ganet: Goal area network for motion forecasting. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1609–1615. IEEE, 2023.

[40] Junqing Wei, John M Dolan, Jarrod M Snider, and Bakhtiar Litkouhi. A point-based mdp for robust single-lane autonomous driving behavior under uncertainties. In *2011 IEEE international conference on robotics and automation*, pages 2586–2592. IEEE, 2011.

[41] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[42] Wei Zhan, Liting Sun, Di Wang, Haojie Shi, Aubrey Clausse, Maximilian Naumann, Julius Kummerle, Hendrik Konigshof, Christoph Stiller, Arnaud de La Fortelle, et al. Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps. *arXiv preprint arXiv:1910.03088*, 2019.

[43] Hang Zhao, Jiyang Gao, Tian Lan, Chen Sun, Ben Sapp, Balakrishnan Varadarajan, Yue Shen, Yi Shen, Yuning Chai, Cordelia Schmid, et al. Tnt: Target-driven trajectory prediction. In *Conference on Robot Learning*, pages 895–904. PMLR, 2021.