REALISTIC WORLD MODEL FOR AUTONOMOUS DRIVING: INTEGRATING PHYSICAL CONSTRAINTS AND MULTI-AGENT INTERACTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Ensuring safety in autonomous driving, particularly in complex and dynamic environments, remains a significant challenge. To address this issue, we propose a novel traffic world model. While existing trajectory forecasting methods typically focus on predicting individual agents and may neglect critical factors such as vehicle dimensions, orientation, and physical constraints, our model incorporates these elements comprehensively. Unlike previous methods that often result in unrealistic scenarios such as collisions or off-road driving, our model integrates physical constraints and introduces innovative loss functions—including safe distance loss and road departure loss—to ensure that the generated trajectories are both realistic and feasible. By simultaneously predicting the trajectories of all agents and explicitly modeling interactions across various scenarios, our approach significantly enhances realism and safety. Our world model functions as a generator, simulator, and trajectory forecasting tool, demonstrating substantial improvements over traditional methods and achieving competitive performance in reducing collision and off-road rates.

1 INTRODUCTION

The past decade has witnessed remarkable progress in autonomous driving, driven by integrating advanced sensors, deep learning algorithms, large language models, and extensive driving datasets. These advancements have led to the development of various autonomous driving systems capable of addressing complex tasks such as perception in adverse weather conditions Li et al. (2023), automatic parking, risk prediction, and end-to-end autonomous driving. Figure1 illustrates end-to-end and modular pipelines used in autonomous driving systems Huang et al. (2023); Gu et al. (2023). Regardless of the pipeline structure, whether end-to-end or modular, safety remains the top priority in autonomous driving.

While end-to-end learning systems have demonstrated the potential for handling diverse driving scenarios, modular pipelines or hybrid approaches continue to be the predominant solutions in the field. This is primarily due to safety concerns, the reliability required for complex driving conditions, and the need for interpretability. As shown in figure1, modular pipelines allow for the decomposition of tasks, such as perception Yang et al. (2023), prediction Feng et al. (2024), and planning Dauner et al. (2023), into distinct components, making it easier to certify, validate, and debug individual subsystems. Most prediction models and planning models do not rely on row data as input; instead, they use a bird's-eye-view (BEV) map that presents road information more clearly and concisely.

Ensuring safety during motion planning is one of the most critical challenges in autonomous driving, particularly in dynamic environments where interactions between multiple agents, such as vehicles, are complex and unpredictable. Existing trajectory forecasting methods primarily focus on predicting the behavior of a single target vehicle while neglecting the broader multi-agent interactions that are essential for realistic and safe driving. Furthermore, many current models are trained on datasets that only contain safe driving scenarios, leading to a lack of robustness when encountering dangerous or high-risk situations. However, collecting these data in the real world is also dangerous and costly.

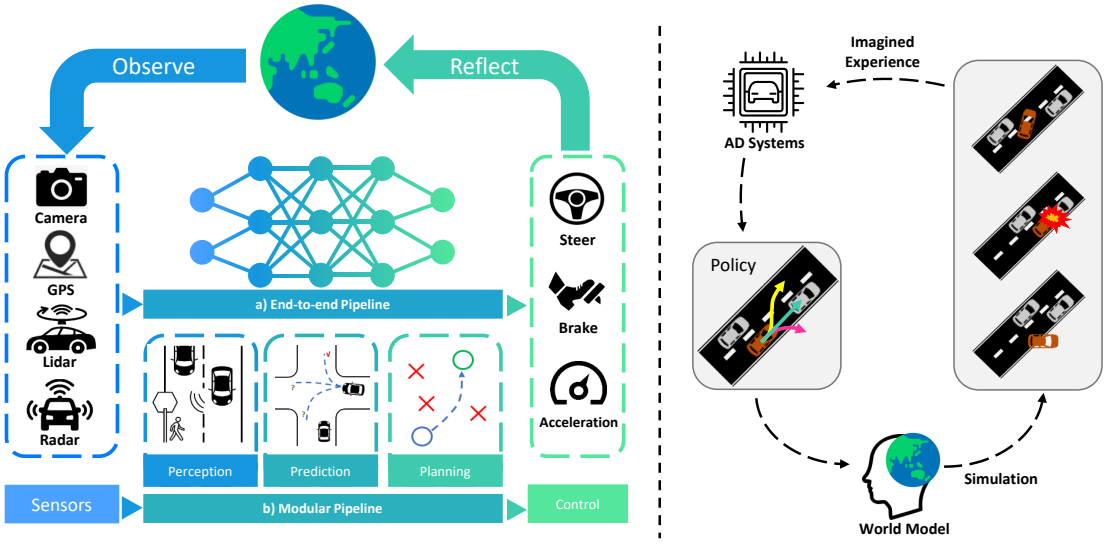

Figure 1: Left: End-to-end pipeline and Modular pipeline for autonomous driving systems. Right: Closed-loop training of autonomous driving systems utilizing the world model.

In this paper, we introduce a novel world model for autonomous driving that directly addresses these limitations by focusing on the interactions between agents within the driving scenes. Our world model simultaneously predicts the future control signals, e.g. steering and acceleration, of all agents in the scene, offering flexibility in whether the prediction is conditioned on the actions of the selected (ego) vehicle or left unconditioned. By modeling multi-agent interactions, the system is able to produce more realistic, dynamic behavior predictions that capture the mutual influence between vehicles. As shown in figure1, it can also enable the closed-loop training Zhang et al. (2022a) of autonomous driving agents.

This interaction-aware approach is crucial for improving the safety of autonomous driving, as agents' behaviors are rarely independent in real-world scenarios. For instance, the movement of one vehicle significantly impacts the trajectories of surrounding vehicles, especially in complex maneuvers such as lane changes, merging, or navigating through intersections. A model that fails to account for these interactions risks generating unsafe or unrealistic trajectories. Our world model explicitly captures these behaviors and ensures that all agent predictions are informed by the collective dynamics of the environment.

In addition, our model introduces two key mechanisms to enhance safety: a safe distance loss (SDL) that discourages collisions between vehicles and a road departure loss (RDL) that ensures adherence to road boundaries. These losses are designed to better handle safety-critical situations, allowing the model to learn from potentially dangerous scenarios. Unlike traditional methods that focus solely on metrics like average displacement error (ADE) or final displacement error (FDE), our approach incorporates these safety constraints to generate more reliable predictions, ultimately reducing collision and off-road incidents.

Our primary contributions include: (1) Proposing a world model that predicts the future trajectories of all agents in the scene, considering both conditioned and unconditioned scenarios based on the ego vehicle's actions; (2) Modeling multi-agent interactions, enhancing the realism and safety of the predictions; (3) Introducing safe distance loss and road departure loss to explicitly address safety-critical scenarios and improve the robustness of the predictions. Our world model addresses the complexities of agent interactions in autonomous driving, representing a significant step toward a safer and more reliable simulation environment rather than just a trajectory forecasting model.

2 RELATED WORKS

2.1 WORLD MODELS FOR AUTONOMOUS DRIVING

The world model LeCun (2022); Ha & Schmidhuber (2018) learns a general representation of the environment and predicts future states based on sequences of actions and the current state. In autonomous driving, such models predict future driving scenarios, such as the behavior of all agents in a scene or driving environment, allowing the system to simulate interactions and make safer decisions. These models can operate on various types of data, including sensor data, such as images and pointclouds, and annotated data, like bird's-eye-view HD maps. An excellent world model not only enhances the safety and reliability of autonomous driving but also forms the foundation for decision-making and planning, enabling effective operation in diverse and dynamic traffic scenarios Cui et al. (2024). Current world models primarily focus on driving scene generation and motion planning on world models.

Driving scene generation entails collecting and processing environmental data from multiple sensors such as LiDAR, cameras, and radar to identify elements like roads, vehicles, pedestrians, and obstacles, thereby constructing accurate environmental models. For instance, GAIA-1 Hu et al. (2023) is a generative world model that leverages video, text, and action inputs to generate realistic driving scenarios while offering fine-grained control over ego-vehicle behavior and scene features. It casts world modeling as an unsupervised sequence modeling problem by mapping the inputs to discrete tokens and predicting the next token in the sequence. DriveDreamer Wang et al. (2023), also dedicated to driving scenario generation, differs from GAIA-1 as it is trained on the nuScenes dataset Caesar et al. (2020). Its model inputs include more elements like HD Maps and 3D boxes, allowing for more precise control over driving scenario generation and deeper understanding, thus improving video generation quality. Additionally, DriveDreamer can generate future driving actions and corresponding predictive scenarios, aiding in decision-making. WorldDreamer Wang et al. (2024a) frames world modeling as an unsupervised visual sequence modeling challenge. This is achieved by mapping visual inputs to discrete tokens and predicting the masked ones. DriveDreamer-2 Zhao et al. (2024) is built upon the framework of DriveDreamer and incorporates a Large Language Model (LLM) to generate user-defined driving videos. The LLM interface is utilized to convert a user's query into agent trajectories, and based on the trajectories, they further generate HD maps.

Motion planning on world models utilizes the world models to determine safe and efficient driving routes. For instance, OccWorld Zheng et al. (2023) and Think2Drive Li et al. (2024) directly utilize 3D occupancy as inputs to predict the evolution of the surrounding environment and plan the actions of autonomous vehicles. Drive-WM Wang et al. (2024b) is a multi-view world model for enhancing the safety of end-to-end autonomous driving planning. Drive-WM, through multi-view and temporal modeling, jointly generates multi-view videos and then predicts intermediate views from adjacent ones, significantly improving consistency across multiple views. SLEDGE Chitta et al. (2024) is a generative simulator for vehicle motion planning trained on real-world driving logs. Its core component is a learned model that is able to generate agent bounding boxes and lane graphs. The model's outputs serve as an initial state for traffic simulation. TrafficBots Zhang et al. (2023) is a multi-agent policy learned from motion prediction datasets. Based on the shared, vectorized context and the individual personality and destination, It can generate realistic multi-agent behaviors in dense urban scenarios. Besides the simulation, TrafficBots can also be used for motion prediction.

In this work, we focus on the physical characteristics of autonomous vehicles. Our model predicts the control signals for all vehicles simultaneously, including steering angles and acceleration, based on the provided BEV HD map. By utilizing the bicycle model and incorporating the actual physical constraints of vehicles, we generate a more realistic autonomous driving world model that aligns closely with real-world physics. This approach improves reliability and realism by focusing on agent interactions while maintaining consistency with the static environment. Moreover, our model is adaptable across different autonomous driving datasets, offering a plug-and-play feature.

2.2 TRAJECTORY PREDICTION

Predicting vehicle trajectories involves inferring future states from observed traffic data and the behavior of all vehicles. Serving as a vital link between perception and planning models, it is essential for safe motion planning. Extensive research in this field has provided solutions from

diverse methods and perspectives Singh (2023); Teeti et al. (2022). The primary methods can be broadly categorized into four types: (1) Physics-based methods, e.g., single trajectory Lin et al. (2000), Monte Carlo Broadhurst et al. (2005) and Kalman filtering Schulz et al. (2018); (2) Classic machine learning methods, e.g., Classic machine learning methods, e.g., Gaussian process Kim et al. (2011), dynamic Bayesian network Jiang et al. (2022); (3) Deep learning methods, for instance, convolutional NN (CNN) Cui et al. (2019), generative model Ivanovic et al. (2020), graph neural networks (GNN) Sheng et al. (2022); Grimm et al. (2023); Pourkeshavarz et al. (2023); Rowe et al. (2023); Zeng et al. (2021); Deo et al. (2022) and transformer-based models Aydemir et al. (2023); Fang et al. (2023); Jiang et al. (2023); Nayakanti et al. (2023); Seff et al. (2023); Kim et al. (2021); Zhou et al. (2023); (4) Reinforcement learning methods, e.g., inverse reinforcement learning Xu et al. (2023); Alsaleh & Sayed (2020); Deo & Trivedi (2020), and generative adversarial imitation learning Zhang et al. (2022b).

3 METHOD

We first built a world model that simulates the driving environment by predicting the future states of all agents in the scene, considering their interactions and the selected vehicle's actions. Our world model incorporates the bicycle model and real-world physical constraints, enabling it to generate realistic driving scenarios that adhere to the physical limitations of vehicles. In Section 3.2, we will explain our overall model, followed by a detailed discussion of the encoder and decoder in Sections 3.3 and 3.4. Our model is significantly inspired by Laformer Liu et al. (2024), utilizing their proposed dense lane-aware prediction framework to enhance the accuracy of trajectory predictions.

3.1 PROBLEM DEFINITION

In the domain of autonomous driving, accurately predicting the future trajectories of surrounding agents is crucial for ensuring safe and efficient navigation. In our work, we consider the world model as a trajectory prediction model that predicts the future trajectories of all agents in the scene simultaneously. Trajectory prediction refers to estimating the future positions of dynamic agents (such as vehicles) based on their current states and the scene context. This involves predicting the future paths of each agent while considering the interactions between them. Our world model operates on a given driving scenario, where the environment remains static, and only the trajectories of the agents are predicted. By simultaneously predicting all agents' trajectories, the model accounts for multi-agent interactions, which is critical for understanding the overall scene dynamics.

The prediction task is framed within the context of a comprehensive HD Map, encompassing detailed representations of lane segments L_{i_L}, where $i_L \in \{0, \cdots, n_L\}$ and historical trajectories of surrounding agents up to the current time step A_{i_A}, where $i_A \in \{0, \cdots, n_A\}$. Additionally, the current actions $\mathbf{c}_{i_A}$ of each vehicle could be considered, including its coordinates, velocity, acceleration, yaw, and yaw rate.

Formally, the lane segments L_{i_L} represent the coordinates of the lane centerlines. We divided the long-range lane centerlines into smaller snippets of a fixed length and discretized them into N_L smaller segments. Our objective is to predict the actions $\mathbf{a}_{i_A}$ of agent i_A over a defined prediction horizon H. Each action $\mathbf{a}_{i_A}$ consists of (a_t, ψ_t), representing the acceleration and steering angle of agent i_A at time step t within the prediction horizon H.

One of the key benefits of this approach is its low cost, as it focuses purely on predicting trajectories without the need to modify or simulate the environment. This makes it an efficient solution for autonomous driving systems, allowing for real-time decision-making and planning based on reliable future state predictions.

3.2 MODEL STRUCTURE

As illustrated in figure 2, our approach utilizes vectorized input to efficiently embed each agent's historical data. This method involves a multi-step process comprising a Local Encoder, a Global Encoder, a Global Local Interaction, and a Decoder. The vectorized input may reduce data size compared to the original image input, thereby enhancing computational efficiency.

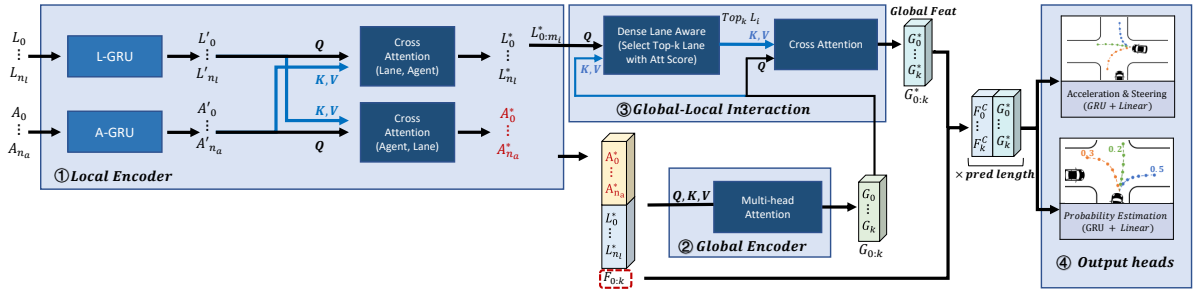

Figure 2: **Structure of Proposed World Model**

The process begins by receiving information from agents and lane segments, vectorized to represent relevant features such as positions, velocities, accelerations, yaws, yaw rates, and lane attributes. The Local Encoder processes these vectorized inputs to extract features specific to each agent and lane segment, capturing temporal dependencies and local interactions within the immediate environment of each entity. The details of the Local Encoder will be explained in Section 3.3.

Following the local encoding, the extracted features, denoted as F_i, are fed into the Global Encoder. The Global Encoder models the global interactions among all agents and lane segments by multi-head attention. This step is crucial for understanding the broader context and interactions within the driving environment, enabling the model to account for complex dependencies and spatial relationships.

Finally, the processed features from the Global Encoder are input into the Decoder. The Decoder is responsible for predicting the next state of each agent, specifically their future control signals, i.e. acceleration and steering. By leveraging the comprehensive representations obtained from both local and global encoders, the Decoder generates accurate predictions of the agents' future actions over the prediction horizon. Afterwards, we can calculate the specific trajectories of the agents through the model output. The details of the Decoder will be explained in Section 3.4.

The original image input was replaced with vectorized input to reduce data size and computational complexity. Vectorized input allows for a more compact representation of the necessary information, facilitating efficient processing and feature extraction while maintaining the integrity of the spatial and temporal data essential for accurate prediction.

3.3 Local Encoder

As illustrated in figure 6, the Local Encoder is designed to effectively process and integrate the historical data of each agent and lane segment information. Let $A_{0:n_a}$ denote the history of each agent and $L_{0:n_l}$ represent the lane segment information.

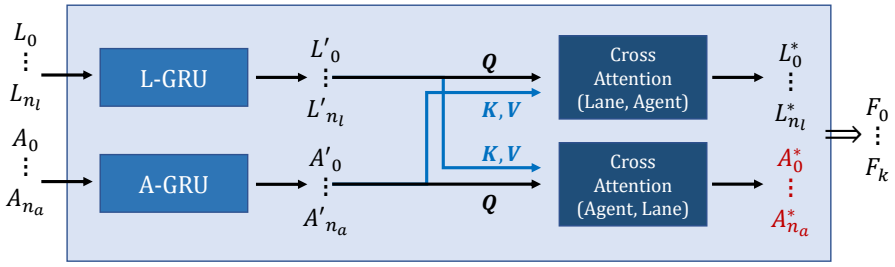

Figure 3: **Structure of Local Encoder.** Here, A_i denotes agents' history trajectories; L_i denotes lane segments.

The inputs $A_{0:n_a}$ and $L_{0:n_l}$ are first embedded via a Gated Recurrent Unit (GRU), which captures temporal dependencies within the historical trajectories of the agents and the static features of the

lane segments. This initial embedding step transforms the raw inputs into latent representations for more effective processing by subsequent layers.

Next, a cross-attention mechanism is employed to obtain contextual information from both agents and lane segments. This cross-attention step allows the model to focus on relevant features and interactions between agents and lane segments, enhancing the local representation of each entity. The output of this step is denoted as $A^*_{0:n_a}$.

By performing cross-attention in both directions on the two inputs, the Local Encoder effectively captures the intricate relationships between agents and lane segments, providing a robust foundation for predicting future trajectories in dynamic driving environments.

3.4 DECODER

As shown in figure 4, the Decoder in our model combines both local and global features to predict the probabilities and the trajectory for the next time step. In this part, we utilized Dense Lane Aware Prediction Liu et al. (2024) to predict possible future destinations, thereby assisting in action prediction. Specifically, the Decoder first integrates lane segment featuresL^* and global features G_i to get dense lane-aware global feat. Then, the dense lane-aware global features and the conditioned local features are utilized to make predictions.

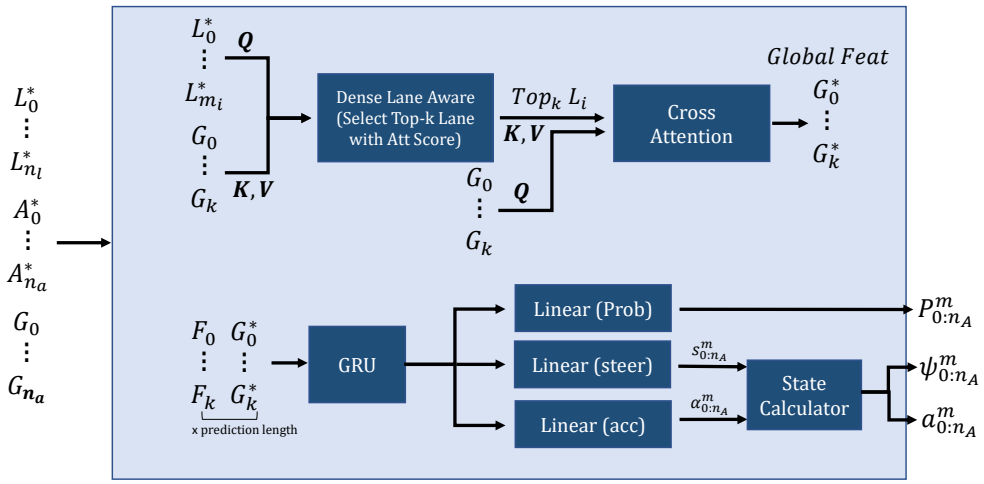

Figure 4: **Structure of Local Encoder.** G_i denotes the global feature from multi-head attention; F_i denotes the local feature obtained from the Local Encoder.

The dense lane-aware model utilizes the global features G_i to predict which lane segment the vehicle is most likely to approach. Attention scores are computed for each lane segment, and from these scores, the top k most likely lane segments are selected. These selected lane segments are then aggregated using a cross-attention mechanism with the global features, effectively enhancing the selected lane segments with the global context.

This process ensures that the model selectively attends to the most probable lane segments based on the global feature G_i, explicitly incorporating the global feature information for these selected lane segments. This selective attention mechanism allows the model to focus on the most relevant parts of the lane structure, improving the accuracy and relevance of the predictions.

After this selective attention process, the local features $F_k = \{A^*_{0:n_a}, L^*_{0:n_l}\}$ are concatenated with the newly refined global features G^*_k. This concatenated representation is then passed through a linear layer to obtain the probabilities of different outcomes. Additionally, the concatenated features are fed into a GRU to predict the future action of the vehicle.

By selectively attending to the most likely lane segments rather than considering the entire lane segment, the Decoder efficiently combines local and global contextual information.

To account for physical limitations, intermediate values after final linear layer $s^m_{0:N_A}$ and $\alpha^m_{0:N_A}$ are both passed through a hyperbolic tangent (tanh) activation function and scaled by predefined limits. Specifically, the acceleration limit a_{lim} is set to 10. The steering angle is also passed through a tanh function and scaled by a steering limit, which decreases as vehicle speed increases. We define steering limit as $s_{lim} = min(0.65, \frac{20}{v^2})$, where v is the vehicle's speed.

Based on the bicycle model, the vehicle's final location is computed using the acceleration and steering angle from the model. For brevity, the detailed derivations and equations are provided in the Appendix A.

This method guarantees that the predicted trajectory adheres to physical constraints, such as steering limits and realistic acceleration thresholds. Additionally, it improves the model's capacity to generate precise, context-aware predictions for subsequent time steps, resulting in more reliable and robust trajectory forecasting in dynamic driving scenarios.

4 EXPERIMENTS

In this section, we present the evaluation of our proposed world model for autonomous driving and the trajectory generator. We conduct extensive experiments using a well-known dataset nuScenes Caesar et al. (2020), showcasing the effectiveness of our approach in predicting future actions and ensuring safe and efficient driving behavior. Since our world model functions similarly to trajectory prediction, and no existing work aligns perfectly with our experimental setup, we compare it with the current trajectory prediction models. The results are analyzed in terms of minADE, minFDE, collision rate, and off-road rate, demonstrating the advantages of our proposed models.

4.1 EXPERIMENT SETUP

In our experiment setting, we have access to large-scale HD Maps, serving as the foundational environment for our world model. This map provides detailed information about lane segments, road structures, intersections, and other critical elements crucial for autonomous driving scenarios. Leveraging this rich dataset, our world model simulates and predicts the movements and interactions of various agents within the driving environment. In our study, we explore three distinct settings to evaluate the performance of the world model in predicting future trajectories of surrounding agents and vehicles in autonomous driving scenarios. Figure 5 shows sample frames of the results of our world model.

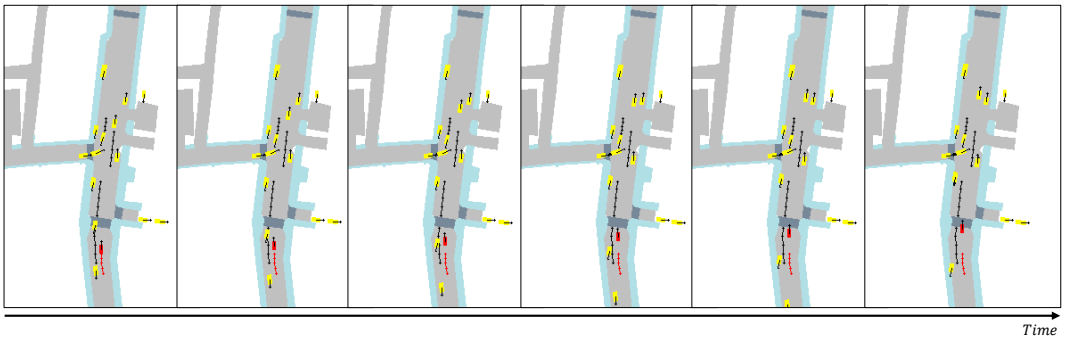

Figure 5: **A sample of the world model.** The red and yellow rectangles are the target vehicle and surrounding vehicles; the black lines are the history trajectories of each agent.

Dataset: For the evaluation of our proposed models, we utilize the nuScenes dataset Caesar et al. (2020), a comprehensive and widely-used dataset in autonomous driving research. The nuScenes dataset offers a rich set of sensory data collected from a fleet of autonomous vehicles operating in diverse urban environments. This dataset contains 1000 scenes every 20 seconds, with ground truth annotations and HD Maps. Vehicles have manually annotated 3D bounding boxes, which are published at 2 Hz.

Metrics: To evaluate our model, we use the standard metrics on the nuScenes leaderboard Caesar et al.. Minimum Average Displacement Error over k ($minADE_k$): The average of pointwise $L2$ distances between the predicted trajectory and ground truth over the k most likely predictions; Minimum Final Displacement Error over k ($minFDE_k$): The final displacement error (FDE) is the $L2$ distance between the final points of the prediction and ground truth. We take the minimum FDE over the k, which is the most likely prediction and average over all agents. To evaluate the safety of our world model, we use two metrics: the collision rate (CR) and the off-road rate (OR). The collision rate measures the frequency of collisions between agents, while the off-road rate quantifies the extent to which agents deviate from drivable areas.

4.2 Conditioned & Unconditioned on target vehicle's action

Table 1: The results of world model conditioned & unconditioned on target vehicle's action.

2 step	$conditioned$	$unconditioned$
$minADE_1$	0.44	0.58
$minFDE_1$	0.60	0.79

We compare two variants of the world model: one conditioned on the target vehicle's action and the other unconditioned. The conditioned variant incorporates predictions of the target vehicle's future trajectory as part of its input. The unconditioned variant directly predicts the target vehicle's action, letting the target vehicle navigate the environment freely. Our experiments indicate that the conditioned approach outperforms the unconditioned model.

4.3 Long-term & Autoregressive prediction

Table 2: The results of world model with long-term & autoregressive short-term prediction.

6 step	$Autoregressive(3 \times 2step)$	$long-term$
$minADE_1$	1.32	1.31
$minFDE_1$	2.61	2.50

We investigate two prediction strategies within the world model framework: long-term prediction and autoregressive short-term prediction. Long-term prediction involves forecasting trajectories over extended time horizons, while autoregressive short-term prediction focuses on immediate future movements. Our results demonstrate that the long-term prediction strategy slightly outperforms the autoregressive approach in accuracy and robustness, emphasizing its effectiveness in capturing complex driving scenarios.

4.4 Ablation Studies

Table 3: **Ablation studies on different output contexts**

output	loss				minADE	minFDE
	l2	arctan	scale	$va\omega r$		
trajectory	✓				3.21	5.83
trajectory	✓	✓			2.71	4.89
trajectory	✓		✓		2.60	4.68
trajectory	✓	✓	✓		2.25	3.97
trajectory+$va\omega r$	✓				2.09	3.68
trajectory+$va\omega r$	✓			✓	**2.07**	**3.66**

In this setting, we conduct an ablation study to examine the impact of different output contexts on the world model's performance. Initially, integrating additional loss functions such as velocity v,

acceleration a, yaw ω, and yaw rate r refined trajectory predictions, significantly enhancing prediction accuracy. Subsequently, we explored directly predicting velocity, acceleration, yaw, and yaw rate outputs but faced challenges in effectively modeling the interactions between these outputs.

5 CONCLUSION

In this study, we have presented a framework for advancing autonomous driving technology, focusing on the world model on High-Definition (HD) Map data. Our experimental evaluations, conducted using the nuScenes dataset, underscore the effectiveness and challenges of each component in achieving safe and efficient autonomous navigation.

The world model developed in this study exhibited promising capabilities in predicting every agent's future actions using a graph-structured representation of the HD Map. By incorporating local and global features, our model achieved enhanced predictive accuracy and computational efficiency across diverse driving scenarios. However, ongoing refinement efforts are necessary to address challenges in modeling nuanced interactions and optimizing output content such as velocity, acceleration, yaw, and yaw rate.

Looking forward, integrating our world model with planning models represents a critical step toward enhancing decision-making processes in real-time driving scenarios. By leveraging predictive insights from the world model alongside planning strategies, we aim to optimize autonomous driving behaviors while ensuring safety and efficiency.

REFERENCES

Rushdi Alsaleh and Tarek Sayed. Modeling pedestrian-cyclist interactions in shared space using inverse reinforcement learning. *Transportation research part F: traffic psychology and behaviour*, 70:37–57, 2020.

Görkay Aydemir, Adil Kaan Akan, and Fatma Güney. Adapt: Efficient multi-agent trajectory prediction with adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8295–8305, 2023.

Adrian Broadhurst, Simon Baker, and Takeo Kanade. Monte carlo road safety reasoning. In *IEEE Proceedings. Intelligent Vehicles Symposium, 2005.*, pp. 319–324. IEEE, 2005.

Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes prediction challenge. `https://www.nuscenes.org/prediction`.

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*, pp. 11621–11631, 2020.

Kashyap Chitta, Daniel Dauner, and Andreas Geiger. Sledge: Synthesizing simulation environments for driving agents with generative models. *arXiv preprint arXiv:2403.17933*, 2024.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 958–979, 2024.

Henggang Cui, Vladan Radosavljevic, Fang-Chieh Chou, Tsung-Han Lin, Thi Nguyen, Tzu-Kuo Huang, Jeff Schneider, and Nemanja Djuric. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *2019 international conference on robotics and automation (icra)*, pp. 2090–2096. IEEE, 2019.

Daniel Dauner, Marcel Hallgarten, Andreas Geiger, and Kashyap Chitta. Parting with misconceptions about learning-based vehicle motion planning. In *Conference on Robot Learning*, pp. 1268–1281. PMLR, 2023.

Nachiket Deo and Mohan M Trivedi. Trajectory forecasts in unknown environments conditioned on grid-based plans. *arXiv preprint arXiv:2001.00735*, 2020.

Nachiket Deo, Eric Wolff, and Oscar Beijbom. Multimodal trajectory prediction conditioned on lane-graph traversals. In *Conference on Robot Learning*, pp. 203–212. PMLR, 2022.

Shaoheng Fang, Zi Wang, Yiqi Zhong, Junhao Ge, and Siheng Chen. Tbp-former: Learning temporal bird's-eye-view pyramid for joint perception and prediction in vision-centric autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1368–1378, 2023.

Lan Feng, Mohammadhossein Bahari, Kaouther Messaoud Ben Amor, Éloi Zablocki, Matthieu Cord, and Alexandre Alahi. Unitraj: A unified framework for scalable vehicle trajectory prediction. *arXiv preprint arXiv:2403.15098*, 2024.

Daniel Grimm, Philip Schörner, Moritz Dreßler, and J-Marius Zöllner. Holistic graph-based motion prediction. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2965–2972. IEEE, 2023.

Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, and Hang Zhao. Vip3d: End-to-end visual trajectory prediction via 3d agent queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5496–5506, 2023.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Yu Huang, Yue Chen, and Zhu Li. Applications of large scale foundation models for autonomous driving. *arXiv preprint arXiv:2311.12144*, 2023.

Boris Ivanovic, Karen Leung, Edward Schmerling, and Marco Pavone. Multimodal deep generative models for trajectory prediction: A conditional variational autoencoder approach. *IEEE Robotics and Automation Letters*, 6(2):295–302, 2020.

Chiyu Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, Dragomir Anguelov, et al. Motiondiffuser: Controllable multi-agent motion prediction using diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9644–9653, 2023.

Yuande Jiang, Bing Zhu, Shun Yang, Jian Zhao, and Weiwen Deng. Vehicle trajectory prediction considering driver uncertainty and vehicle dynamics based on dynamic bayesian network. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(2):689–703, 2022.

ByeoungDo Kim, Seong Hyeon Park, Seokhwan Lee, Elbek Khoshimjonov, Dongsuk Kum, Junsoo Kim, Jeong Soo Kim, and Jun Won Choi. Lapred: Lane-aware prediction of multi-modal future trajectories of dynamic agents. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14636–14645, 2021.

Kihwan Kim, Dongryeol Lee, and Irfan Essa. Gaussian process regression flow for analysis of motion trajectories. In *2011 International Conference on Computer Vision*, pp. 1164–1171. IEEE, 2011.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Jinlong Li, Runsheng Xu, Jin Ma, Qin Zou, Jiaqi Ma, and Hongkai Yu. Domain adaptive object detection for autonomous driving under foggy weather. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 612–622, 2023.

Qifeng Li, Xiaosong Jia, Shaobo Wang, and Junchi Yan. Think2drive: Efficient reinforcement learning by thinking in latent world model for quasi-realistic autonomous driving (in carla-v2). *arXiv preprint arXiv:2402.16720*, 2024.

Chiu-Feng Lin, A Galip Ulsoy, and David J LeBlanc. Vehicle dynamics and external disturbance estimation for vehicle path prediction. *IEEE Transactions on Control Systems Technology*, 8(3): 508–518, 2000.

Mengmeng Liu, Hao Cheng, Lin Chen, Hellward Broszio, Jiangtao Li, Runjiang Zhao, Monika Sester, and Michael Ying Yang. Laformer: Trajectory prediction for autonomous driving with lane-aware scene constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2039–2049, 2024.

Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratarth Goel, Khaled S Refaat, and Benjamin Sapp. Wayformer: Motion forecasting via simple & efficient attention networks. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2980–2987. IEEE, 2023.

Mozhgan Pourkeshavarz, Changhe Chen, and Amir Rasouli. Learn tarot with mentor: A meta-learned self-supervised approach for trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8384–8393, 2023.

Luke Rowe, Martin Ethier, Eli-Henry Dykhne, and Krzysztof Czarnecki. Fjmp: Factorized joint multi-agent motion prediction over learned directed acyclic interaction graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13745–13755, 2023.

Jens Schulz, Constantin Hubmann, Julian Löchner, and Darius Burschka. Multiple model unscented kalman filtering in dynamic bayesian networks for intention estimation and trajectory prediction. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1467–1474. IEEE, 2018.

Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S Refaat, Rami Al-Rfou, and Benjamin Sapp. Motionlm: Multi-agent motion forecasting as language modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8579–8590, 2023.

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.

Apoorv Singh. Trajectory-prediction with vision: A survey. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3318–3323, 2023.

Izzeddin Teeti, Salman Khan, Ajmal Shahbaz, Andrew Bradley, Fabio Cuzzolin, and Lud De Raedt. Vision-based intention and trajectory prediction in autonomous vehicles: A survey. In *IJCAI*, pp. 5630–5637, 2022.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023.

Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. Worlddreamer: Towards general world models for video generation via predicting masked tokens. *arXiv preprint arXiv:2401.09985*, 2024a.

Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14749–14759, 2024b.

Yifan Xu, Theodor Chakhachiro, Tribhi Kathuria, and Maani Ghaffari. Solo t-dirl: Socially-aware dynamic local planner based on trajectory-ranked deep inverse reinforcement learning. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12045–12051. IEEE, 2023.

Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17830–17839, 2023.

Wenyuan Zeng, Ming Liang, Renjie Liao, and Raquel Urtasun. Lanercnn: Distributed representations for graph-centric motion forecasting. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 532–539. IEEE, 2021.

Chris Zhang, Runsheng Guo, Wenyuan Zeng, Yuwen Xiong, Binbin Dai, Rui Hu, Mengye Ren, and Raquel Urtasun. Rethinking closed-loop training for autonomous driving. In *European Conference on Computer Vision*, pp. 264–282. Springer, 2022a.

Qichao Zhang, Yinfeng Gao, Yikang Zhang, Youtian Guo, Dawei Ding, Yunpeng Wang, Peng Sun, and Dongbin Zhao. Trajgen: Generating realistic and diverse trajectories with reactive and feasible agent behaviors for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(12):24474–24487, 2022b.

Zhejun Zhang, Alexander Liniger, Dengxin Dai, Fisher Yu, and Luc Van Gool. Trafficbots: Towards world models for autonomous driving simulation and motion prediction. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1522–1529. IEEE, 2023.

Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv preprint arXiv:2403.06845*, 2024.

Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. *arXiv preprint arXiv:2311.16038*, 2023.

Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17863–17873, 2023.

A APPENDIX

A.1 CALCULATING LOCATIONS USING SIMPLIFIED BICYCLE MODEL

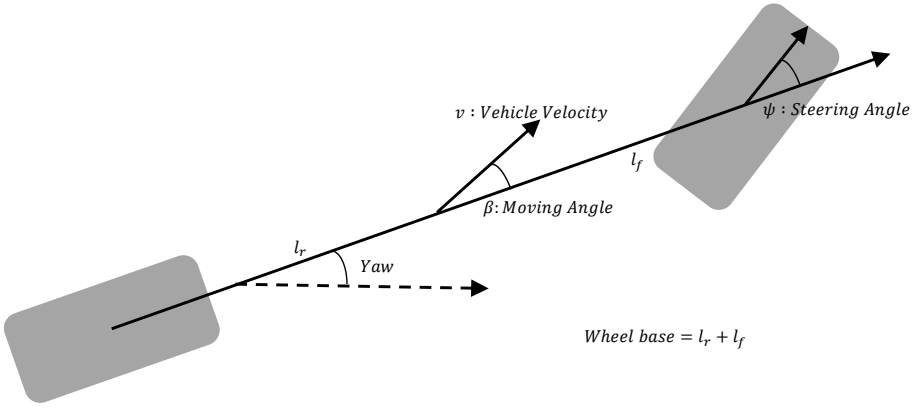

Figure 6: **Illustration of Bicycle Model**

The next location and yaw of the vehicle are computed using the acceleration and steering angle as follows. The model outputs the calculated location and yaw as the final result.

1. Next velocity is determined by:

$$v_{\text{next}} = v_{\text{current}} + \Delta t \cdot a_t$$

2. The moving angle is determined by:

$$\beta = \arctan\left(\frac{\psi}{2}\right)$$

3. The change in yaw is computed as:

$$\Delta\text{yaw} = \frac{v_{current}}{\text{wheel\_base}} \cdot \tan(\psi) \cdot \cos(\beta)$$

4. The updated yaw is:

$$\text{yaw}_{\text{next}} = \text{yaw}_{\text{current}} + \Delta\text{yaw} \cdot \Delta t$$

5. The moving direction is:

$$\theta = \text{yaw}_{\text{current}} + \beta$$

6. The displacement (action) is computed as:

$$\text{action} = v \cdot \Delta t \cdot (\cos(\theta), \sin(\theta)))$$

7. Finally, the vehicle's next location is:

$$\text{loc}_{\text{next}} = \text{loc}_{\text{current}} + \text{action} \cdot \Delta t$$