Diffusion-Based Simulation Agents for Scalable Autonomous Driving

Jianwei Ren

Abstract

Autonomous driving is revolutionizing traditional transportation, yet constructing simulators to validate its algorithms remains highly challenging. This proposal formulates the task as joint distribution modeling of multi-agent states, with the goal of aligning the distribution with real-world driving scenarios. By achieving this alignment, the simulator can freely sample multi-modal trajectories for dynamic evolution. Inspired by recent advancements in generative foundation models like Sora [1], this proposal sketches diffusion models as the backbone of a stochastic virtual world to underpin agents' behavior logic. This approach promises to substantially reduce the cost of algorithm validation, enhance efficiency, and address safety concerns.

1. Introduction

The advancement of autonomous driving technology promises to transform transportation by enhancing safety, efficiency, and accessibility. However, despite its potential, the field faces significant challenges that impede its development and widespread adoption.

Key bottlenecks in the development of autonomous vehicles (AVs) include the stochastic nature of real-world driving environments, the complexity of agent interactions, and the need for rigorous safety validation. These challenges create a pressing need for innovative testing methodologies that can replicate the intricacies of real-world scenarios. Traditional methods, like playback testing—where real-world sensor data is replayed with minor software adjustments to observe potential outcomes—fall short due to their inability to account for adaptive responses from other vehicles and road users.

In this context, simulation agents (sim agents) emerge as a crucial solution. Sim agents are AI-driven entities designed to replicate human behaviors and driving patterns within simulated environments. By utilizing sim agents, researchers and developers can create diverse, realistic, and scalable scenarios to test autonomous driving systems in a cost-efficient manner. Unlike traditional log-playback methods, which often overestimate aggressiveness by sticking rigidly to planned routes, or rule-based methods that are overly accommodating and reactive, sim agents strike a balance by accurately representing the full distribution of human behavior. Integrating them into the development process not only improves the reliability of AV algorithms by ensuring they can handle a variety of dynamic conditions, but also aids in safety validation by identifying potential risks before real-world deployment, minimizing the likelihood of accidents and system failures.

Sim agents play a pivotal role in the development of autonomous driving systems, yet the field remains relatively underdeveloped. Recently, Waymo introduced a benchmark [8] that has spurred further interest in the field. In this proposal, I will review existing approaches to sim agents development and suggest a potential solution to address the challenges involved.

2. Related Work

2.1. Multi-Agent Traffic Simulation

Sim agents task frames simulation as a distribution matching problem, requiring the simulator to closely match the distribution of human driving behaviors to achieve authenticity. It shares similarities with motion prediction and planning, but differs in nuanced ways in objectives, outputs, and constraints. Simulation agent modeling employs closed-loop evaluation [8], ensuring realism in behaviors and dynamic interactions. Although both sim agents and motion prediction can forecast trajectories in multi-agent scenarios, the latter is primarily open-loop and focuses on marginal predictions, limiting its ability to recover from out-of-domain situations. Compared to planning, sim agent modeling addresses a more general problem, as each agent can execute a replica of a planner independently.

2.2. Multi-Modal Motion Prediction

Motion prediction is probabilistic and multi-modal in nature, reflecting the stochasticity of the real world: even under same initial conditions, agents can exhibit diverse yet patterned behaviors and trajectories. To faithfully predict an unbiased distribution of possible futures, generative models like diffusion [17] and autoregressive models [14] have gained popularity for their exceptional ability to

model complex distributions. Traditional regression models mold the output space to parametric continuous distributions, such as mixtures of Gaussian [10] or Laplace [18], leveraging heuristic and explicit motion goal candidates [4] or queries [12] to achieve multi-modality. In contrast, generative models freely sample from the learned space, thereby better reflecting the inherent boundlessness and uncertainty of the real world.

Autoregressive models [9, 14, 15, 19] have demonstrated strong performance, but still suffer from error accumulation. Moreover, predicting the next discrete token does not fully align with this task, as the action space for trajectories is continuous, resulting in inevitable degradation. In contrast, diffusion models [5, 6, 13, 16, 17] offer an alternative solution. Matching the scalability and interactive capabilities of aforementioned GPTs, they excel in modeling continuous trajectories. Even more promising is their ability to avoid causal confusion and enable controllable generation through conditions.

Notably, although motion prediction typically focuses on a subset rather than all agents, as sim agents do, it still suggests potential for extension to the latter task, as evidenced by [10] and [15].

3. Method

3.1. Problem Formulation

The objective of simulation is to generate possible scenarios by sampling from a world model that captures the interactions between agents and their context. In this proposal, the world of the simulator is formulated as $P(\mathcal{S}, \mathcal{O})$, where $\mathcal{S}$ represents the state of agents and $\mathcal{O}$ denotes the environment. Due to resource limitations, the simulator operates within a finite spatial and temporal range. Without loss of generality, we define the "current" time as $t = 0$, at which the scenario is observed.

The context at each timestep $o_t \sim \mathcal{O}$ can be decomposed into static components $o_t^s = o_0$, such as lanes, and dynamic components o_t^d, such as traffic signals. By sampling in a well-defined initialization space under conditional guidance c, the joint distribution of "future" agent states is presented as:

$$p(s^1_{>0}, ..., s^k_{>0} | s^1_{\leq 0}, ..., s^k_{\leq 0}, o^d_{\leq 0}, o^s_0, c) \tag{1}$$

where $s^K \sim \mathcal{S}$ represents the states of K observable agents within a finite spatial domain. $s^K_{\leq 0}, o^d_{\leq 0}, o^s_0$ can be readily obtained from real-world datasets collected in various driving scenarios. c is a composite condition encompassing predefined spatiotemporal dimensions and manually specified constraints, such as traffic regulations.

Specifically, when this distribution is applied to marginal agents, these agents are treated as sim agents. If trajectories are further obtained through sampling strategies, the task transitions into a planning problem.

3.2. Diffusion the World Model

Recent studies [1, 2, 11] have shown that diffusion models basically capture physical laws, interaction logic, and game rules. Additionally, advancements [7] in embodied AI highlight the diffusion transformer's (DiT) strengths in data integration and scalability, inspiring this proposal to adopt such a generative foundation model as the backbone for the virtual world.

Concretely, diffusion progressively adds Gaussian noise to a scenario (or its latent embedding), creating a sequence of increasingly noisy states. This forward process, modeled as a Markov chain, ensures the final state becomes a nearly isotropic Gaussian distribution. A model is trained to predict and denoise these noisy states by minimizing the discrepancy between the denoised output and the original agent states at each step. During inference, the model iteratively denoises a randomly sampled Gaussian noise, refining it into a plausible and realistic sequence in the continuous space. Moreover, conditions can be incorporated throughout the process to generate trajectories that meet specific requirements or constraints.

The description above reveals four reasons why diffusion is well-suited for autonomous driving simulation: their ability to model continuous spaces, achieve multi-modality through sampling, capture the joint distribution of multiple agents, and, last but no least, incorporate specific conditions like navigation and intention to guide generation.

3.3. Dataset and Evaluation

Dataset. The dataset for the sim agents challenge is based on the large-scale Waymo Open Motion Dataset (WOMD) [3], comprising over 531K scenarios for training and validation. Each scenario includes map details, such as lanes, crosswalks, boundaries, and traffic signals, as well as history-future trajectory pairs for vehicles, pedestrians, and cyclists. Each complete trajectory consists of a 9 second 10 Hz sequence, capturing the key states of interest for the simulator: position and heading.

Evaluation. As described in [8], realism is assessed by measuring the alignment between the distributions of simulated and real agents. Since the distribution lacks an analytic form, this high-dimensional objective is approximated through various measurements, including kinematic, object interaction, and map-based metrics. The metric is calculated as a convex combination over the components. This evaluation is closed-loop and forward-looking, facilitating long-term planning and simulation.

4. Conclusion

Simulation agents represent a critical tool for addressing the challenges in AV development, enabling realistic and scalable testing environments. By leveraging diffusion models

as the generative backbone, this proposal outlines a plausible path to achieve enhanced simulation accuracy and adaptability. This proposal seeks to accelerate progress within the AV community by enabling cost-effective and efficient algorithm validation.

References

[1] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1, 2

[2] Decart, Julian Quevedo, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. 2024. 2

[3] 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. 2

[4] Junru Gu, Chen Sun, and Hang Zhao. Densetnt: End-to-end trajectory prediction from dense goal sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15303–15312, 2021. 2

[5] Zhiming Guo, Xing Gao, Jianlan Zhou, Xinyu Cai, and Botian Shi. Scenedm: Scene-level multi-agent trajectory generation with consistent diffusion models. *arXiv preprint arXiv:2311.15736*, 2023. 2

[6] 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*, pages 9644–9653, 2023. 2

[7] Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024. 2

[8] Nico Montali, John Lambert, Paul Mougin, Alex Kuefler, Nicholas Rhinehart, Michelle Li, Cole Gulino, Tristan Emrich, Zoey Yang, Shimon Whiteson, et al. The waymo open sim agents challenge. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2

[9] Jonah Philion, Xue Bin Peng, and Sanja Fidler. Trajeglish: Traffic modeling as next-token prediction. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[10] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Mtr++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[11] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024. 2

[12] Balakrishnan Varadarajan, Ahmed Hefny, Avikalp Srivastava, Khaled S Refaat, Nigamaa Nayakanti, Andre Cornman, Kan Chen, Bertrand Douillard, Chi Pang Lam, Dragomir Anguelov, et al. Multipath++: Efficient information fusion and trajectory aggregation for behavior prediction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7814–7821. IEEE, 2022. 2

[13] Junming Wang, Xingyu Zhang, Zebin Xing, Songen Gu, Xiaoyang Guo, Yang Hu, Ziying Song, Qian Zhang, Xiaoxiao Long, and Wei Yin. He-drive: Human-like end-to-end driving with vision language models. *arXiv preprint arXiv:2410.05051*, 2024. 2

[14] Yu Wang, Tiebiao Zhao, and Fan Yi. Multiverse transformer: 1st place solution for waymo open sim agents challenge 2023. *arXiv preprint arXiv:2306.11868*, 2023. 1, 2

[15] Wei Wu, Xiaoxin Feng, Ziyan Gao, and Yuheng Kan. Smart: Scalable multi-agent real-time simulation via next-token prediction. *arXiv preprint arXiv:2405.15677*, 2024. 2

[16] Ziyuan Zhong, Davis Rempe, Yuxiao Chen, Boris Ivanovic, Yulong Cao, Danfei Xu, Marco Pavone, and Baishakhi Ray. Language-guided traffic simulation via scene-level diffusion. In *Conference on Robot Learning*, pages 144–177. PMLR, 2023. 2

[17] Ziyuan Zhong, Davis Rempe, Danfei Xu, Yuxiao Chen, Sushant Veer, Tong Che, Baishakhi Ray, and Marco Pavone. Guided conditional diffusion for controllable traffic simulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3560–3566. IEEE, 2023. 1, 2

[18] Zikang Zhou, Zihao Wen, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Qcnext: A next-generation framework for joint multi-agent trajectory prediction. *arXiv preprint arXiv:2306.10508*, 2023. 2

[19] Zikang Zhou, Haibo Hu, Xinhong Chen, Jianping Wang, Nan Guan, Kui Wu, Yung-Hui Li, Yu-Kai Huang, and Chun Jason Xue. Behaviorgpt: Smart agent simulation for autonomous driving with next-patch prediction. *arXiv preprint arXiv:2405.17372*, 2024. 2