

Enhancing Autonomous Driving through Dual-Process Learning with Behavior and Reflection Integration

Xiao Zhang

University of Science and Technology Beijing

Kangsheng Wang

University of Science and Technology Beijing <https://orcid.org/0009-0009-8392-4148>

Tianyu Hu

tianyu@ustb.edu.cn

University of Science and Technology Beijing <https://orcid.org/0000-0001-9903-0696>

Huimin Ma

mhmpub@ustb.edu.cn

University of Science and Technology Beijing

Research Article

Keywords: Cognitive, Autonomous Driving, Large Language Model, Dual-Process Theory

Posted Date: March 18th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-6223799/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Enhancing Autonomous Driving through Dual-Process Learning with Behavior and Reflection Integration

Xiao Zhang¹, Kangsheng Wang¹, Tianyu Hu^{1,†}, and Huimin Ma^{1,†}

¹ School of Computer and Communication Engineering,
University of Science and Technology Beijing,
Beijing, China

Abstract—Contemporary autonomous driving (AD) methodologies, which predominantly convert visual features into control directives, face long-tail challenges due to constraints imposed by limited data distribution. Conversely, human drivers exhibit proficiency in such conditions, underscoring the significance of emulating human cognition in AD systems. Therefore, we introduce Dual-Process Learning (D-PL) approach for cognitive-enhanced decision-making. Inspired by dual-process theory, the D-PL method combines Behavior Pattern Learning (BPL) and Self-Reflective Learning (SRL) to integrate quick, intuitive decisions with deliberate, analytical reasoning, constructing a hierarchical decision model for sophisticated trajectory planning. Our approach improves decision-making, enhances adaptability, and tackles the crucial open-world generalization challenge encountered by current AD methods. Comprehensive evaluations on the nuScenes dataset validate the robustness of our method, demonstrating its superior performance in navigating the intricacies of real-world contrasting with conventional models.

Index Terms—Cognitive, Autonomous Driving, Large Language Model, Dual-Process Theory

I. INTRODUCTION

Contemporary autonomous driving (AD) methodologies primarily concentrate on translating visual inputs into control commands [1], [2]. While these methods demonstrate proficiency in controlled environments, their heavy dependence on restricted data distribution impedes their performance in unpredictable open-world scenarios. This dependency restricts adaptability, especially in rare, crucial safety situations referred to as long-tail challenges [3]–[5]. In contrast to the above data-driven approaches, human drivers exhibit the ability to generalize to unseen scenarios through common sense and reasoning skills, highlighting the urgent need to integrate human cognition into AD systems [6].

Drawing inspiration from dual-process theory [7], human intelligence functions on two levels: 1) a behavior pattern learning process that is intuitive, and limited by training data, and 2) a self-reflective reasoning process that is logical, and excels at handling long-tail scenarios through reasoning.

To this end, we present Dual-Process Learning (D-PL), a hierarchical decision method that combines Behavior Pattern Learning (BPL) with Self-Reflective Learning (SRL) to train Large Language Models (LLMs) [8] in acquiring the reasoning pattern from human drivers. LLMs demonstrate proficiency in applying common sense and reasoning, essential for managing long-tail scenarios [9]. BPL, involving pattern establishment and acquisition, empowers the LLM to emulate the behavior patterns of the oracle agent during the mastery of intricate tasks and presupposed strategies. SRL benefits from tokenized language representations derived from the self-reflective mechanism. These representations serve as reward signals for LLMs to self-assess and adapt when encountering unpredictable scenarios.

By integrating BPL and SRL within a dual-process framework, our method embeds human cognition into the AD system, thereby enhancing its ability to generalize in open-world environments. Our experiments on the nuScenes dataset [10] demonstrate the effectiveness of D-PL, highlighting its superior performance compared to traditional models. In conclusion, our contributions are: (1) presenting the D-PL framework, which incorporates human cognition into the AD system based on dual-process theory, (2) verifying the utility of LLMs for intricate decision-making in open-world AD scenarios, and (3) demonstrating the effectiveness of D-PL through comprehensive evaluations on the nuScenes dataset.

II. SEMANTIC CONTROL IN AUTONOMOUS DRIVING

Traditional AD systems [1], [2] often rely on mapping visual features directly to control signals. However, their performance is constrained by the training data distribution [4]. Recent studies have utilized LLMs for trajectory planning by framing it as a language modeling task. LLMs, trained on extensive datasets, exhibit robust capabilities in common-sense reasoning. GPT-Driver [11] applies fine-tuned LLMs for planning, while Agent-Driver [12] employs LLMs as driving agents, demonstrating encouraging outcomes in real-world trials. Nonetheless, these methods simply map queries to responses, lacking the capacity for analytical thought and self-reflection, which hinders their ability to reason effectively

[†]: Corresponding author

This work was supported in part by National Nature Science Foundation of China (No. 62172036) and National Science and Technology Major Project (2022ZD0116305)

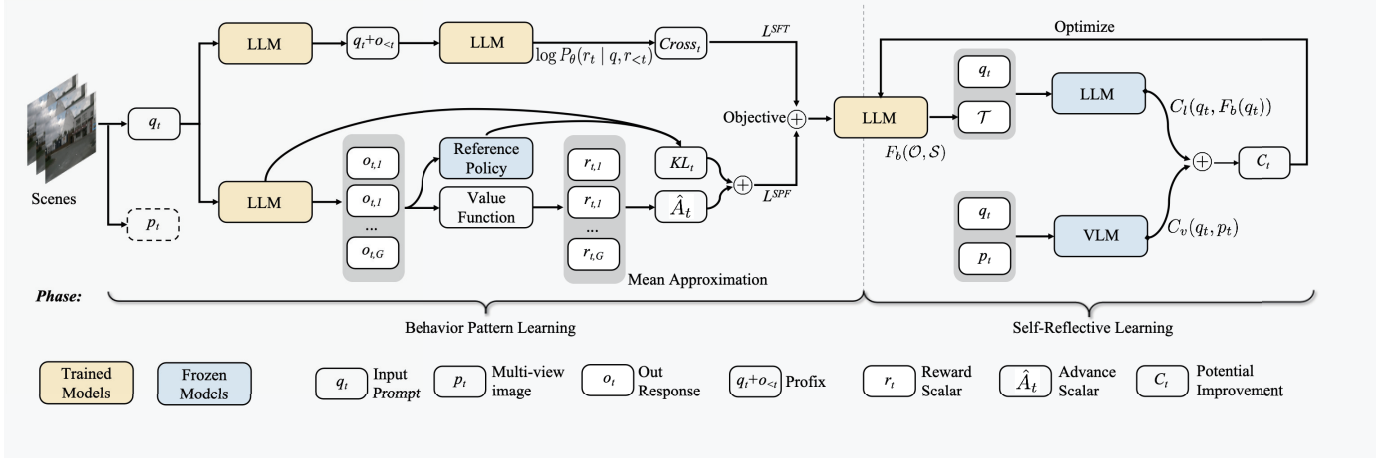


Fig. 1. Hierarchical decision framework of Dual-Process Learning, which integrates Behavior Pattern Learning and Self-Reflective Learning.

in open-world scenarios [6]. Our approach, grounded in dual-process theory, bridges these gaps by amalgamating intuitive and analytical decision-making processes, thereby bolstering robustness and generalization.

III. PROPOSED DUAL-PROCESS FRAMEWORK

In this section, we introduce the D-PL framework, a hierarchical decision method. Illustrated in Figure 1, D-PL operates through two primary stages. The initial stage, BPL, acquires distinct driving behavior patterns by leveraging textual descriptions provided by the oracle agent in the nuScenes dataset. These descriptions, comprising observations $\mathcal{O}$ and ego states $\mathcal{S}$, constitute dataset $\mathcal{D}$, capturing human reasoning patterns. The trajectory planning within BPL is denoted by $F_b(\cdot)$ (refer to Equation 1), where the LLM devises the trajectory by conforming to the behavior pattern of the oracle agent. In the second stage, SRL constructs a self-improvement framework, depicted as $F_s(\cdot)$ (refer to Equation 2), where the LLM replans trajectories based on the task-specific instruction.

$$\mathcal{T} = F_b(\mathcal{O}, \mathcal{S}) \quad (1)$$

$$\mathcal{T}^* = F_s(F_b(\mathcal{O}, \mathcal{S})) \quad (2)$$

Both $F(\cdot)$ functions are formulated as $\text{LLM}(\mathcal{O}, \mathcal{S} | \mathcal{I})$, where $\mathcal{I}$ represents the instructions directing LLM in tasks such as trajectory planning, reflection, and re-planning.

A. Behavior Pattern Learning for Reasoning Basis

Drawing inspiration from Behavior Cloning (BC) [13], we utilizes seq2seq learning [14] to emulate the behavior of the oracle agent, effectively managing fundamental visual-motor tasks while encountering challenges with varied behaviors. We explicitly characterize the behavior patterns of the oracle agent, enabling the LLM to acquire more generalized patterns.

1) *Behavior Pattern Modeling:* Reasoning plays a pivotal role in human decision-making [15]. Table I delineates our driving behavior pattern for reasoning, integrating Chain of Thought (CoT) [16] reasoning into the trajectory planning process. Commencing with $\mathcal{O}$ and $\mathcal{S}$, our model advances

through three stages: Critical Target Analysis (CTA), Behavior Decision-making (BDE), and Trajectory Planning (TPL). In alignment with the above sequential stages, we convert the structured data $\mathcal{O}$ and $\mathcal{S}$ into sequential input prompt q_t that the LLM can process through CoT reasoning.

TABLE I
THE KEY COMPONENTS OF BEHAVIOR PATTERN MODELING

Stage	Definition	Effect
CTA	We use CoT to instigate the LLM to analyze critical targets and latent impacts step-by-step.	Focusing on the key elements of the road and reducing cognitive resources.
BDE	LLM guides compound actions consisting of meta actions like speed maintenance, lane changes, and narrowing choices with CoT.	High-level action decisions aim to reduce the reasoning search space and deliver optimized planning strategies.
TPL	LLM predicts efficient, safe waypoints by integrating situational awareness with behavior decisions in a CoT task.	Crucial module for predicting the vehicle's trajectory with environmental tokens.

2) *Stabilized Policy Fine-tuning:* We leverage Proximal Policy Optimization (PPO) [17] and present a Stabilized Policy Fine-tuning (SPF) to align LLM parameterized to θ with the desired behavior pattern. Nevertheless, using four large LLM models during fine-tuning led us to explore approximations to reduce resource demands. Specifically, we focused on approximating the advantage value $\hat{A}_t$ by averaging reward groups (see Equation 3) [18]. In lieu of conventional PPO rewards, we additionally utilized a task-specific value function $V(q_t)$ predicated on the Euclidean distance between predicted and oracle agent in $\mathcal{D}$. This approach reduced the frequency of policy updates by leveraging globally aggregated information, thereby enhancing both the stability and efficiency of the policy in targeted tasks. We optimize the LLM π_θ using the

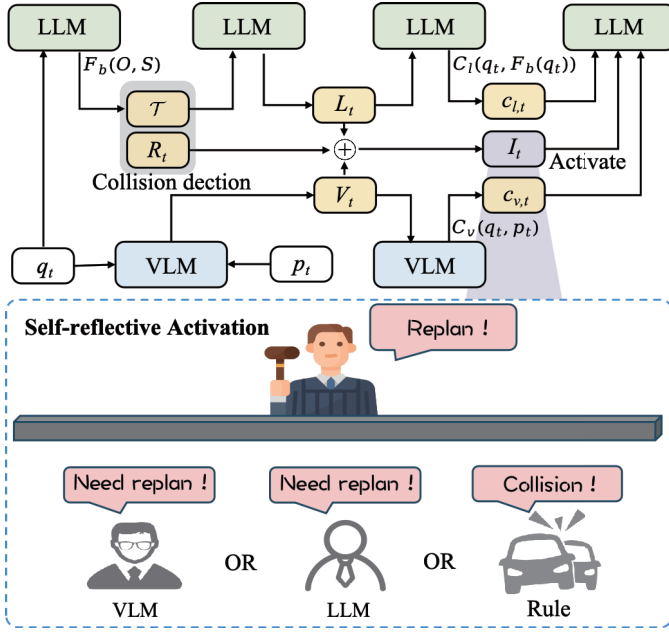


Fig. 2. Block diagram of self-improvement in Self-Reflective Learning.

following objective:

$$L^{SPF}(\theta) = \mathbb{E}_t \left[\frac{1}{G} \sum_{i=1}^G \min \left(r_t(\theta) \hat{A}_t, \right. \right. \\ \left. \left. \text{clip} \left(r_t(\theta), 1 - \epsilon, 1 + \epsilon \right) \hat{A}_t \right) - \beta D_{KL}(\pi_\theta \parallel \pi_{\text{ref}}) \right] \quad (3)$$

In this scenario, $\hat{A}_t$ is computed by averaging, with weights determined by perplexity (PPL). A lower PPL indicates higher model confidence in the sample. $r_t(\theta)$ represents the ratio between the new policy π_θ and the reference policy π_{ref} , an LLM with frozen parameters. Moreover, Supervised Fine-Tuning (SFT) [19] serves as a regularization term to establish behavior patterns from the oracle agent and improve the numerical reasoning of the LLM (refer to Equation 4). Our proposed fine-tuning of SPF is based on a separate SFT fine-tuned LLM, which further ensures SPF stability by ensuring the stability of the outcome format.

$$L^{SFT}(\theta) = - \sum_{(q,o) \in \mathcal{D}} \sum_{t=1}^T \log P_\theta(o_t \mid q, o_{<t}). \quad (4)$$

B. Self-Reflective Learning for Self-Adjustment

While BPL acquires the behavior pattern through BC, it encounters challenges in effectively generalizing and adapting the behavior pattern, crucial capabilities for open-world. Moreover, it is difficult to compensate for perceptual omissions that affect the integrity of the description of the environment [20]. To the end, we proposed the multimodal self-reflective mechanism, illustrated in Figure 2, which amalgamates the reasoning capabilities of the Vision Language Model (VLM) [21], [22] and the LLM through Instruction Learning [8] that

enables a model to follow natural language commands to perform specific tasks.

1) *Self-reflective mechanism*: Concurrently, the LLM processes both language context and predicted trajectory, generating rewards represented by sequential tokens. Reward signals partake in self-optimization, addressing potential shortcomings of planning shown in Equation 5.

$$c_{l,t} = C_l(q_t, F_b(q_t)), \quad (5)$$

Here, C_l denotes the LLM generation task that scrutinizes the planning, detects errors and potential improvements, and offers optimal planning solutions.

2) *Perceptual complement*: The VLM uses cross-attention to integrate multi-view images p_t and query q_t , aligning features between the two modalities. This mechanism enables the VLM to detect and correct omissions in the visual data. Combined with its reasoning capabilities, the VLM provides the LLM with a refined context. The reward signals from the VLM used to trigger LLM re-planning are outlined as follows.

$$c_{v,t} = C_v(q_t, p_t), \quad (6)$$

where C_v effectively parses and resolves visual information to verify the coherence of the environment description and complement it. Furthermore, C_v also scrutinizes the decision recommendations for the present scenario. Consequently, Equation 2 is adjusted as follows.

$$\mathcal{T}_t^* = F_s(F_b(q_t); \text{concat}(c_{l,t}, c_{v,t})). \quad (7)$$

The variables $c_{l,t}$ and $c_{v,t}$ are incorporated into the prompt that directs the LLM to perform re-planning. This multimodal self-reflective mechanism enhanced the integrity of information transmission, thereby augmenting D-PL capability for self-assessment and refinement, which in turn leads to significantly enhanced accuracy and adaptability in long-tail task.

3) *Self-reflective activation*: Following the delineation of the two inference levels in our hierarchical decision-making method, a heuristic algorithm is employed to regulate the activation of self-reflective. BPL is favored for its intuitive reasoning, ensuring precise alignment between context and planning based on past actions. The system transitions to SRL when met the following conditions:

$$I_t = R_t \vee L_t \vee V_t \quad (8)$$

where R_t represents the governing rule for collision detection derived from [2]. L_t and V_t are the evaluations of the current scenario and corresponding reasoning from the LLM and VLM, respectively, reflecting their self-reflective capabilities.

IV. EXPERIMENTAL EVALUATION

A. Dataset and performance metrics

For trajectory planning, the observation $\mathcal{O}$ and ego-state $\mathcal{S}$ are obtained from the nuScenes dataset, a comprehensive urban AD dataset, are transformed into unstructured textual data to be interpreted by the LLM. We analyzed 28,130

manually annotated human driving records presented in a dialogic format. We evaluated D-PL and baseline methods based on safety (Collision rate), the primary metric, as well as personalized metrics such as comfort (Jerk, refer to Equation 9) and energy efficiency (Energy, refer to Equation 10) for advanced driving tasks.

$$\text{Jerk} = \frac{1}{N} \sum_{i=1}^N \|B''(t + \delta_t) - B''(t)\| \quad (9)$$

$$\text{Energy} = \int_{\substack{t \in [0,1] \\ B'(t) \cdot B''(t) > 0}} m B'(t) \cdot B''(t) dt \quad (10)$$

where $B'(\cdot)$ and $B''(\cdot)$ represent the first and second derivatives of the Bézier curve [23] for the trajectory planning results, respectively. N denotes the number of samples along in Bézier curve. The weight m is configured as 1.

TABLE II
SAFETY, COMFORT, AND ENERGY EFFICIENCY SCORES IN GENERAL SCENARIOS.

Metrics	Method	Collision rate (%)↓				Jerk(%)↓	Energy (J)↓
		1s	2s	3s	Avg.		
ST-P3	ST-P3 [24]	0.23	0.62	1.27	0.71	8.04	14472.83
	VAD [25]	0.09	0.12	0.26	0.16	7.96	11375.9
	GPT-Driver [11]	0.04	0.12	0.36	0.17	10.3	12168.79
	Agent-Driver [12]	0.03	0.08	0.56	0.23	8.97	11878.65
	D-PL	0.03	0.09	0.25	0.12	4.30	7856.13
UniAD	FF [26]	0.11	0.35	0.93	0.46	9.36	13728.67
	EO [27]	0.05	0.18	0.93	0.39	13.50	10208.72
	UniAD [2]	0.06	0.21	0.83	0.37	12.45	12179.39
	GPT-Driver	0.04	0.16	0.52	0.24	-	-
	Agent-Driver	0.03	0.11	0.67	0.27	-	-
	D-PL	0.03	0.12	0.37	0.17	-	-
Human		-	-	-	-	6.86	9648.98

B. Comparison with state-of-the-art methods

As depicted in Table II, our approach surpassed both baseline methods in terms of ST-P3 [24] and UniAD [2] metrics, notably reducing the average collision rate by a significant margin. This highlights the effectiveness of D-PL in improving safety and managing diverse tasks. Our method achieved a Jerk of 4.30% and energy consumption of 7856.13 J, surpassing human drivers, who exhibit a Jerk of 6.86% and energy consumption of 9648.98 J, as well as outperforming other baseline methods. The outcomes of above Open-loop testing [28] validate that this performance improvement is derived from optimizing comfort and energy efficiency measures rather than solely conforming to the oracle agent.

TABLE III
GENERALIZATION ABILITY OF OUR METHOD ON TRAINING SETS OF VARIOUS SIZES

Method	UniAD				D-PL			
	1%	10%	50%	100%	0%	1%	10%	100%
Avg.Col (%) ↓	7.12	1.16	0.57	0.28	5.71	2.19	1.32	0.42
Avg.L2 (m) ↓	5.41	1.79	1.07	0.47	3.26	1.51	1.05	0.77
Invalid outputs ↓	-	-	-	-	18	2	0	0

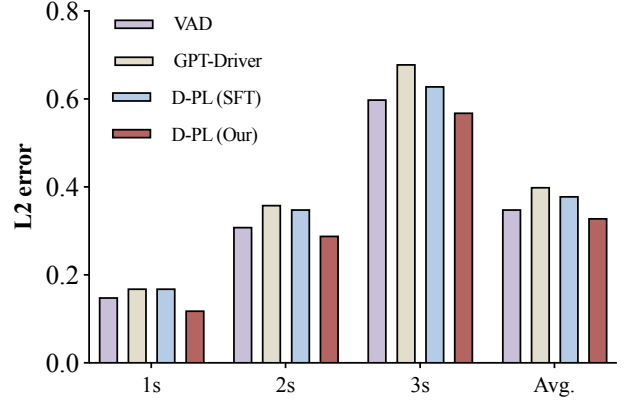


Fig. 3. Bar chart of L2 error compared to the state-of-the-arts.

C. Few-shot Learning

Humans possess the remarkable ability to swiftly acquire proficiency in intricate tasks with minimal exposure, generalizing effectively from limited examples and rapidly adapting with sparse data. This skill is recognized as few-shot learning. Table III distinctly illustrates that our approach excels in the few-shot learning task. This demonstrates that our approach can accomplish swift generalization with solely the limited sample to be navigated. Furthermore, our method also demonstrates the ability of zero-shot setting (0% training sample).

D. Effect of Stabilized Policy Fine-tuning

Figure 3 shows the L2 error between the predicted trajectories and the oracle agent for all planning methods. After applying SPF, our approach achieves state-of-the-art performance with an average L2 error of 0.33. Notably, beyond the figure, ST-P3 yields an average L2 error of 1.86, and our method without fine-tuning records 3.26. These results underscore the stability and precision introduced by SPF, significantly improving performance over baseline methods.

V. CONCLUSION

In this paper, we addressed the limitations of current AD systems, which struggle to handle real-world unpredictability. We proposed the D-PL framework, grounded in dual-process theory, which combines BPL for efficient handling of routine tasks and SRL for deeper reasoning and adaptability in unfamiliar scenarios. By integrating LLMs with dual-process theory, D-PL enhances the AD system's capability to generalize effectively in long-tail situations. Our evaluations on the nuScenes dataset demonstrated the robustness of D-PL, highlight superiority over conventional models. This work provides a crucial step toward embedding human-like cognition in AD systems, with future efforts focused on advancing real-time decision-making to further improve safety and efficiency in dynamic, open-world environments.

REFERENCES

- [1] F. Kunz, D. Nuss, J. Wiest, H. Deusch, S. Reuter, F. Gritschneider, A. Scheel, M. Stubler, M. Bach, P. Hatzelmann et al., "Autonomous driving at ulm university: A modular, robust, and sensor-independent fusion approach," in IEEE intelligent vehicles symposium, 2015, pp. 666–673.
- [2] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang et al., "Planning-oriented autonomous driving" in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 17 853–17 862.
- [3] N. Kshetry and L. R. Varshney, "Safety in the Face of Unknown Unknowns: Algorithm Fusion in Data-driven Engineering Systems," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2019, pp. 8162–8166.
- [4] Y. Zhou, G. Tan and C. Gou, "Hierarchical Home Action Understanding with Implicit and Explicit Prior Knowledge," in 2024 IEEE International Conference on Acoustics, Speech and Signal Processing, 2024, pp. 4015–4019.
- [5] Y. Liu, Y. Li, Z. Liu, W. Yang, Y. Wang and Q. Liao, "Clip-Based Synergistic Knowledge Transfer for text-based Person Retrieval," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2024, pp. 7935–7939.
- [6] B. Y. Lin, Y. Fu, K. Yang, F. Brahman, S. Huang, C. Bhagavatula, P. Ammanabrolu, Y. Choi, and X. Ren, "Swiftsage: A generative agent with fast and slow thinking for complex interactive tasks," Advances in Neural Information Processing Systems, 2024, vol. 36.
- [7] P. C. Wason and J. S. B. Evans, "Dual processes in reasoning?" in Cognition, 1974, vol. 3, no. 2, pp. 141–154.
- [8] J. Wei, M. Bosma, V.-Y. Zhao, K. Guu, et al, "Finetuned language models are zero-shot learners," in International Conference on Learning Representations, 2022.
- [9] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong et al. "The rise and potential of large language model based agents: A survey." arXiv preprint arXiv:2309.07864 (2023).
- [10] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in Computer Vision and Pattern Recognition, 2020, pp. 11 621–11 631.
- [11] J. Mao, Y. Qian, H. Zhao, and Y. Wang, "Gpt-driver: Learning to drive with gpt," arXiv preprint arXiv:2310.01415, 2023.
- [12] J. Mao, J. Ye, Y. Qian, M. Pavone, and Y. Wang, "A language agent for autonomous driving," arXiv preprint arXiv:2311.10813, 2023.
- [13] F. Torabi, G. Warnell, P. Stone, "Behavioral cloning from observation," in International Joint Conference on Artificial Intelligence, 2018, pp. 4950–4957.
- [14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Conference of the North, 2018, pp. 4171–4186.
- [15] J.-S.-B.-T. Evans, D.-E. Over, "Rationality and reasoning," in Psychology Press, 2013.
- [16] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi et al., "Chain-of-thought prompting elicits reasoning in large language models," in Neural Information Processing Systems, 2022, vol. 35, pp. 24 824–24 837.
- [17] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.
- [18] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, M. Zhang, Y. Li, Y. Wu, and D. Guo, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models" arXiv preprint arXiv:2402.03300, 2024.
- [19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez et al, "Attention is all you need, " in Neural Information Processing Systems, 2017, vol. 30.
- [20] Y. Yang, X. Zhang, J. Xu and W. Han, "Empowering Vision-Language Models for Reasoning Ability through Large Language Models," in IEEE International Conference on Acoustics, Speech and Signal Processing, 2024, pp. 10 056–10 060.
- [21] D. H. Ballard, Y. Chen, "A multimodal learning interface for word acquisition," in IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003, pp. V-784.
- [22] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, J. Gao, "Vinvl: Revisiting visual representations in vision-language models" in IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5579–5588.
- [23] W. Boehm and A. Muller, "On de casteljau's algorithm," in Computer Aided Geometric Design, 1999, vol. 16, no. 7, pp. 587–605.
- [24] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning," in European Conference on Computer Vision, 2022, pp. 533–549.
- [25] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, et al, "Vad: Vectorized scene representation for efficient autonomous driving," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 8340–8350.
- [26] P.-Y. Hu, A. Huang, J. Dolan, D. Held, D. Ramanan, "Safe Local Motion Planning with Self-Supervised Freespace Forecasting," in IEEE/CVF Computer Vision and Pattern Recognition Conference, 2021, pp. 12 732–12 741.
- [27] T. Khurana, P.-Y. Hu, A. Dave, J. Ziglar, D. Held, D. Ramanan, "Differentiable Raycasting for Self-supervised Occupancy Forecasting," in European Conference on Computer Vision, 2022, pp. 9675–9684.
- [28] Y. Kang, H. Yin, and C. Berger, "Test yourself-driving algorithm: An overview of publicly available driving datasets and virtual testing environments," IEEE Transactions on Intelligent Vehicles, 2019, vol. 4, no. 2, pp. 171–185.