# Continuously Learning, Adapting, and Improving: A Dual-Process Approach to Autonomous Driving

**Jianbiao Mei**[1,2,*] **Yukai Ma**[1,2,*] **Xuemeng Yang**[2] **Licheng Wen**[2] **Xinyu Cai**[2]
**Xin Li**[2,4] **Daocheng Fu**[2] **Bo Zhang**[2] **Pinlong Cai**[2] **Min Dou**[2]
**Botian Shi**[2,†] **Liang He**[3] **Yong Liu**[1,†] **Yu Qiao**[2]
[1] Zhejiang University [2] Shanghai Artificial Intelligence Laboratory
[3] East China Normal University [4] Shanghai Jiao Tong University

## Abstract

Autonomous driving has advanced significantly due to sensors, machine learning, and artificial intelligence improvements. However, prevailing methods struggle with intricate scenarios and causal relationships, hindering adaptability and interpretability in varied environments. To address the above problems, we introduce **LeapAD**, a novel paradigm for autonomous driving inspired by the human cognitive process. Specifically, LeapAD emulates human attention by selecting critical objects relevant to driving decisions, simplifying environmental interpretation, and mitigating decision-making complexities. Additionally, LeapAD incorporates an innovative dual-process decision-making module, which consists of an Analytic Process (System-II) for thorough analysis and reasoning, along with a Heuristic Process (System-I) for swift and empirical processing. The Analytic Process leverages its logical reasoning to accumulate linguistic driving experience, which is then transferred to the Heuristic Process by supervised fine-tuning. Through reflection mechanisms and a growing memory bank, LeapAD continuously improves itself from past mistakes in a closed-loop environment. Closed-loop testing in CARLA shows that LeapAD outperforms all methods relying solely on camera input, requiring 1-2 orders of magnitude less labeled data. Experiments also demonstrate that as the memory bank expands, the Heuristic Process with only 1.8B parameters can inherit the knowledge from a GPT-4 powered Analytic Process and achieve continuous performance improvement. Project page: https://pjlab-adg.github.io/LeapAD/.

## 1 Introduction

Since the early 21st century, starting with the DARPA Grand Challenge [1], humanity has explored replacing human drivers with computer algorithms. Over the past two decades, advancements in sensor technology, machine learning, and artificial intelligence have propelled the evolution of self-driving technology. Recent data-driven approaches achieved considerable success, as evidenced by new vehicle models featuring intelligent driving assistance and the commercial operation of L4 robotaxis in several cities [2–4]. However, these methods depend heavily on diverse training data distributions, resulting in a superficial understanding of underlying semantics and potential misconceptions in complex situations. This is because data-driven approaches primarily perform induction on observed patterns without the capability for deduction, thus constraining their performance to the coverage of the annotated data. Therefore, there is an urgent need for a system capable of reasoning about unseen scenarios and utilizing knowledge in a human cognition manner.

* equal contribution, † corresponding author

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

The latest advancements in Large Language Models (LLMs) and Vision Language Models (VLMs), noted for their embedded world knowledge and robust explanatory and reasoning capabilities, have captured the interest of researchers [5–8]. For example, in the autonomous driving field, some knowledge-based methods [9–12] employ LLMs and VLMs as the driving agents. However, these methods perform open-loop testing, which merely evaluates errors between model output and the ground truth from datasets, failing to reflect the dynamic interactions between ego car and the real-world environment [13]. Consequently, they are often inadequate to effectively assess the responsiveness and adaptability of driving agents.

In fact, human learning to drive involves a continuous interaction and exploration process within closed-loop environments, where drivers make decisions based on the surroundings and receive feedback accordingly. As per the dual-process theory [14–16], human intelligence operates on two levels: 1) *Heuristic Process* (*System-I*), which is automatic, quick, empirical, and domain-specific; and 2) *Analytic Process* (*System-II*), which is rational, slow, and excels in logical reasoning and creativity across various domains. This dual-process thinking is evident in the progression from novice to experienced driver. Initially, individuals rely heavily on common sense due to their lack of driving experience. Through training, they develop driving skills via a closed-loop learning process involving continuous trial and error, along with rational analysis (Analytic Process) to evaluate their behavior. These skills become internalized over time, forming muscle memory that enables quick, instinctive reactions in familiar driving scenarios (Heuristic Process). Even after obtaining driver's license, individuals continue to gain experience and learn from accidents to enhance driving skills.

To this end, we develop a dual-process closed-loop autonomous driving system that is continuously **le**arning, **a**dapting and im**p**roving, named **LeapAD**. Similar to the human attention mechanism, the scene understanding module in LeapAD mainly focuses on critical objects that may affect driving decisions, simplifying the environmental description and the decision-making process. Following such scene understanding, we develop a dual-process decision-making module that emulates human cognitive processes, featuring a Heuristic Process and an Analytic Process [17, 15]. Through a closed-loop setup, Analytic Process accumulates experience and builds a transferable memory bank of high-quality driving decisions. The knowledge can be adapted to various scenarios and then transferred to the lightweight model in Heuristic Process through supervised fine-tuning (SFT). The Heuristic Process is employed for closed-loop decision-making using a few-shot strategy. When traffic accidents occur, the Analytic Process intervenes to analyze these incidents and update the memory bank, enabling the system to continuously improve through self-reflection. The main contributions of our work are summarized as follows:

- We develop an innovative closed-loop autonomous driving approach that emulates the critical object attention mechanisms and the learning processes observed in human driving behavior.
- We propose a dual-process decision-making module inspired by human cognition theory. In the absence of human involvement, our approach enables the fast, empirical Heuristic Process to inherit the capabilities of the slow, rational Analytic Process in a self-supervised manner.
- LeapAD utilizes the Analytic Process and a reflection mechanism to accumulate a transferable memory bank, enabling the system to achieve continuous learning and generalization capabilities in a closed-loop driving environment.
- Extensive experiments in CARLA show that LeapAD not only outperforms all other methods relying solely on camera input, but also achieves this with 1-2 orders of magnitude less annotated data.

## 2 Related Works

### 2.1 Large Vision Language Models

Inspired by the successful deployment of Large Language Models (LLMs) like LLaMAs [18, 19] and Vicuna [20], a plethora of Vision Language Models (VLMs) [21–29] has emerged to broaden their applicability to multi-modal understanding. Various models, such as BLIP2 [30] which utilizes the Q-former, Flamingo [21] leveraging a perceiver resampler, and LLaVA [31] alongside MiniGPT-4 [32] that incorporate instruction tuning, have been innovated to enhance feature alignment, few-shot learning, and create versatile visual agents. Moreover, models like Qwen-VL [28] with its three-stage training, and InternVL [29]'s image-text alignment method, assist in achieving advanced multi-lingual and fine-grained visual comprehension. The rise of VLM and visual-language-action (VLA) models has injected the vitality of autonomous driving, presenting researchers with new opportunities.

### 2.2 Empowering Autonomous Driving with Foundation Models

Recent work [33, 34, 9–11, 35] explores the use of large foundation models in autonomous driving, leveraging their embedded world knowledge and powerful interpretation and reasoning capabilities. For understanding driving scenarios, a series of datasets and benchmarks [36, 37, 6, 38] have been proposed. To improve the interpretability of autonomous driving, LMDrive [9] and DriveMLM [39] use LLMs to generate human-instructed decisions in the simulated environment, which is data-dependent and hard to adapt to the real world. Agent-Driver [10] adopts LLM agent for planning, which is less efficient due to excess environmental data. ELM [8] introduces a vision-language model tailored for embodied understanding within driving scenarios. RAG-Driver [11] improves driving interpretation and signal prediction by integrating retrieval augmentation and in-context learning. Recent DriveVLM-Dual [12] integrates VLM and data-driven planning pipelines, providing solutions for deployment. Contrary to the mentioned techniques, our LeapAD draws from attention mechanisms and observational learning and decision-making in human driving. It utilizes a memory bank for experience storage and replay in a closed-loop scenario, enabling continuous learning via memory and reflection mechanisms.

### 2.3 From Data-Driven to Knowledge-driven Autonomous Driving

While the prevailing data-driven approaches [2–4, 40–49] have led to success in both academia and industry in past decades, allowing autonomous driving technology to be used in people's daily lives. However, these methods are limited to the distribution of training data and frequently encounter adaptability issues and long-tail challenges when expanding across different areas [50, 51]. On the other hand, human drivers possess a deep common sense of understanding the world, which enables them to adapt to unexpected scenarios. This highlights the need for a shift to the knowledge-driven approaches, which involve using empirical reasoning and induction to learn from the environment [52, 53], and updating insights to develop specialized skills [54, 55]. Knowledge-driven methods acquire general knowledge rather than merely implementing predefined human rules or abstracting characteristics from collected data in specific domains [7]. These approaches enhance performance, interpretability, and safety by integrating human-like logic into AI systems, particularly in managing complex traffic scenarios. In the era of foundation models, the advanced reasoning and knowledge application capabilities exhibited by LLMs and VLMs have proven highly effective for complex tasks such as understanding, reasoning, and decision-making within the domain of autonomous driving [33, 56, 57]. These foundation models have embedded world knowledge and robust explanatory and reasoning capabilities through extensive training on diverse datasets and captured the researchers' interest [5–8].

## 3 Methodology

### 3.1 Overview

In this section, we introduce how we design our anthropomorphic closed-loop autonomous driving system, LeapAD. Figure 1 illustrates that LeapAD consists of three main components: the VLM for scene understanding (Section 3.2), the dual-process decision-making module comprising the Analytic Process (Section 3.3) and the Heuristic Process (Section 3.4), along with the action executor for low-level control (Appendix A). In the CARLA simulator, LeapAD utilizes VLM to process the surrounding images and generate descriptions of critical objects. These scene descriptions are then fed into the dual-process decision-making module in order to derive scene reasoning and driving decisions. Finally, these high-level decisions are forwarded to the action executor, translated into control signals, and interact with the simulator.

In closed-loop driving environments, the fine-tuned lightweight model in Heuristic Process is used to perform quick, empirical decisions with the transferable experience in the memory bank. And when the Heuristic Process encounters accidents, the Analytic Process intervenes. The Analytic Process exploits LLMs to analyze traffic accidents, leveraging its embedded world knowledge, particularly its understanding of traffic rules. It then generates corrected, high-quality driving experiences, enriching the memory bank and enabling continuous learning for the entire system.

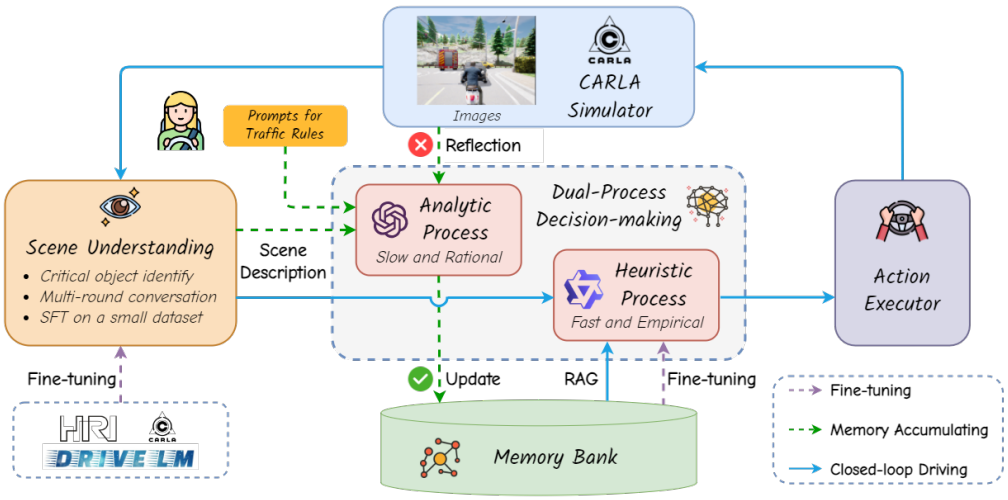



Figure 1: The detailed architecture of our proposed LeapAD. The scene understanding module analyzes surrounding images and provides descriptions of critical objects that may influence driving decisions. These scenario descriptions are then fed into the dual-process decision module, which drives reasoning and decision-making. The generated decisions are then transmitted to action executors, where they are converted into control signals for interaction with the simulator. The Analytic Process then uses an LLM to accumulate experience in driving analysis and decision-making, conducting reflections on accidents. The experience is stored in the memory bank and transferred to a lightweight language model, forming our Heuristic Process for quick responses and continuous learning.

### 3.2 Scene Understanding with VLM

Human drivers typically focus on critical elements surrounding the vehicle to prevent information overload, enhance reaction time, and minimize cognitive load. This approach helps improve driving concentration and reduces accident probabilities. Inspired by such a mechanism, the scene understanding module in LeapAD is designed to selectively identify critical objects, simplifying the description of the surrounding environment and reducing the load on decision-making processes.

Specifically, since off-the-shelf foundation VLMs lack domain-specific knowledge in the driving domain, we perform SFT and prompt the VLMs to output the linguistic descriptions of the objects that may influence subsequent driving decisions. The description of these critical objects includes their semantic, spatial, motion attributes, and behavioral reasoning. Integrating these aspects promotes a comprehensive understanding of the environment, which can ensure safety and adaptability in complex and dynamic driving environments. For a specific driving scene, the descriptions generated by VLM can be expressed as $D = \{A_{s,i}, A_{l,i}, A_{m,i}, C_{r,i}\}_{i=0}^{N-1}$, where $N$ denotes the number of the critical objects. For each critical object $O_i$, the description contains: *i)* the semantic attribute $A_s$ describes its semantic category, usually important traffic participants (e.g., vehicles and cyclists) and infrastructure (e.g., traffic lights and stop signs). *ii)* The spatial attribute $A_l$ indicates its bounding box, the lane it locates, and the distance from the ego car, which are important for safety and collision avoidance. *iii)* The motion attribute $A_m$ refers to the motion direction of the object. *iv)* Behavioral reasoning $C_r$ describes why the object is critical and how it influences the driving decision of the ego car. For example, when the ego car goes straight, the stop sign on the right side is of high importance because it indicates the need to stop at the intersection. We provide an example to further illustrate the descriptions of critical objects in the driving scene, as shown in Figure 9 in Appendix B. Notably, the VLM not only excels in simulated environments but also demonstrates robust performance in real-world scenarios.

### 3.3 Analytic Process

Based on the scene descriptions provided by the VLM, we design the Analytic Process to imitate the rational thinking of a human driver. The Analytic Process relies on logical reasoning, employing rational thinking to analyze complex situations and make safe driving decisions. The LLMs, through their extensive pre-training on diverse datasets, have encapsulated vast amounts of world knowledge, equipping them with the ability to handle intricate problems with nuanced understanding

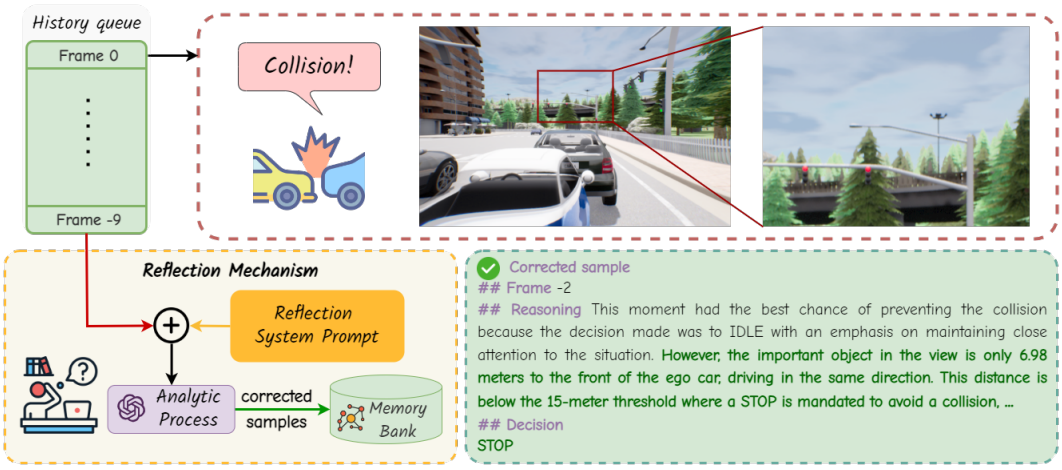



Figure 2: Detailed procedure of the reflection mechanism. When Heuristic Process encounters traffic accidents, the Analytic Process intervenes, analyzing historical frames to pinpoint errors and provide corrected samples. These corrected samples are then integrated into the memory bank to facilitate continuous learning.

and reasoning [34]. This capability aligns with the requirements of the Analytic Process in driving scenarios, where decisions must be made based on deep analysis and contextual understanding of the environment. Our Analytic Process harnesses the power of LLMs, leveraging its world knowledge to understand the scene descriptions and perform high-quality driving analysis and decisions. We empirically found that prompting LLMs with specific traffic rules provided in Appendix C further improves safety and is more reliable for on-road scenarios.

Furthermore, we integrate the VLM and the Analytic Process to run closed-loop experiments and collect the high-quality decision-making processes and results generated by the Analytic Process as "experience" in a memory bank. The accumulated experience can be seamlessly transferred to the Heuristic Process, facilitating it to react quickly based on experience when handling similar situations, as described in Section 3.4.

**Reflection mechanism.** We also employ the Analytic Process to reflect on traffic accidents, as shown in Figure 2. Specifically, when VLM and Heuristic Process run in a closed-loop driving scenario, any accident will trigger the reflection mechanism. During this procedure, the scene description $D$, reasoning $R$, and decision $S$ of the preceding frames before the accident are forwarded to Analytic Process. It is then required to meticulously analyze the cause of the event, locate the error, and provide corrected reasoning and decisions. The insights gained from the reflection procedure are further integrated into the memory bank, allowing the LeapAD to continuously learn from failures and progressively lead to more informed and accurate decision-making in future driving scenarios. Importantly, the experience in the memory bank has good transferability and generalization. It can be directly utilized by other lightweight models and easily generalized to different scenarios, as demonstrated in section 4.4.

### 3.4 Heuristic Process

While the Analytic Process can offer more precise driving reasoning and decisions due to its detailed analysis and careful consideration, the inherent slow processing causes duplicated and redundant effort, limiting its application in practical driving scenarios. In contrast, human drivers form muscle memory through repeated practice and experience, requiring less effort over time. To reflect this quick and empirical thinking pattern and facilitate practical application, we craft a Heuristic Process in LeapAD incorporating a lightweight language model. Specifically, we perform supervised fine-tuning (SFT) using the samples stored in the accumulated memory bank mentioned in Section 3.3 to distill knowledge into the lightweight language model. By this means, the Heuristic Process achieves behavior adaption to various scenarios and runs much faster than Analytic Process (about 5 times faster in our experiments). We empirically found that the lightweight model without SFT is unable to produce appropriate driving decisions.

**Few-shot Prompting.** Moreover, we perform few-shot prompting [34] to enhance the Heuristic Process's generalization ability for unseen scenes and mitigate hallucinations for more robust decisions.

Through such a mechanism, the Heuristic Process can effectively leverage the experience and deep insights from the existing memory bank, improving the accuracy of future driving decisions. To facilitate the retrieval of similar driving scenes from the memory bank for few-shot prompting, we primarily rely on the embedding similarity between the current scene's descriptions and those stored in the memory bank. However, if directly calculating text similarity based on the original descriptions, the presence of redundant linguistic information in descriptions can complicate the differentiation between scenes. Thus, we propose a novel scene encoding method to extract and encode compressed captions that comprise key elements such as the critical object's category, lane, and distance from the ego car. This approach streamlines the procedure of querying similar scenarios, enhancing retrieval efficiency and accuracy by prioritizing the most influential aspects for driving decisions of each scenario. Afterward, the compressed captions are sent to a text encoder to encode the embedding vectors, which are expressed as follows:

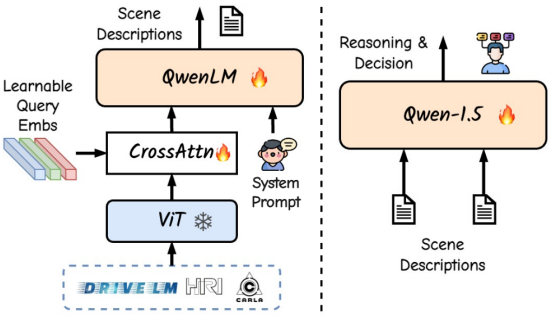



Figure 3: The illustration of the fine-tuning process. We fine-tune the VLM (Qwen-VL-7B) using 11K instruction-following data for scene understanding (left). Also, we utilize the collected samples in the memory bank to fine-tune Qwen-1.5 used in Heuristic Process, as illustrated in the right part.

$$\mathbf{e} = T_e(F_c(D)), \tag{1}$$

where $D$ is the scene description, $F_c$ denotes the compressing process, and $T_e$ indicates the text encoder. Subsequently, the cosine similarity between the query embedding $\mathbf{e}_q$ for the current scene and the embedding $\{\mathbf{e}_i\}_{i=0}^{M-1}$ for the memory bank with the size of $M$ is computed by:

$$s(\mathbf{e}_q, \mathbf{e}_i) = \frac{\mathbf{e}_q \cdot \mathbf{e}_i}{\|\mathbf{e}_q\| \|\mathbf{e}_i\|}. \tag{2}$$

We select top-k samples with the highest similarity scores as queried scenes. The scene descriptions $\{D_i\}_{i=0}^{k-1}$, reasoning $\{R_i\}_{i=0}^{k-1}$, and decisions $\{S_i\}_{i=0}^{k-1}$ of k samples and the scene descriptions $D_c$ of the current scene are both fed into the Heuristic Process for the final reasoning $R_c$ and decision $S_c$.

# 4 Experiments

## 4.1 Data preparation

**Data for VLM.** We construct the instruct-following datasets for supervised fine-tuning of our VLM by integrating Rank2Tell [58], DriveLM [36], and data collected within CARLA [59]. To maintain consistency across all the datasets, we adopt a uniform standard reference format for critical objects as: `<ref>In {camera view}, {properties}</ref><box>{coordinates}</box>`. For each dataset, specific Q&A pairs are created to suit their unique structures and contents. The conversations are structured in a summary-elaboration manner, with the first question for the Rank2Tell and DriveLM datasets focusing on determining the number, semantic, and spatial attributes, such as the bounding box coordinates of key objects. For the Rank2Tell dataset, we follow up by inquiring about the moving state, importance, and corresponding reasoning for each critical object. For the DriveLM dataset, we retain most of the original questions but eliminate redundant ones. We extract only 6K frames of data from Rank2tell and DriveLM and organize them in the standardized format. The data gathered from the CARLA simulator is exclusively dedicated to the closed-loop experiments detailed in Section 4.3. A comprehensive training dataset of 5K frames is collected from Town 01-04, 06, 07, and 10. To identify key objects within the scene, we design several automatic annotation rules, as detailed in Appendix B. For clarity, we provide the data illustration in Figure 9 in Appendix B.

**Data for Heuristic Process.** We leverage the integration of Analytic Process and VLM to accumulate experience within the closed-loop setup and save it in the memory bank for subsequent SFT and few-shot prompting of Heuristic Process. Moreover, our approach incorporates dynamic updates to address errors encountered by the Heuristic Process, as mentioned in the reflection mechanism

Table 1: Comparison of our LeapAD with competitive methods on Town05 Short benchmark. Notably, LeapAD demonstrated superior performance with a smaller data footprint, outperforming other approaches. "DD" & "KD" denote data-driven and knowledge-driven, respectively. "L" & "C" indicate LiDAR and camera modalities.

| Method | Modality | Type | Annotations | DS ↑ | RC ↑ |
|---|---|---|---|---|---|
| InterFuser [41] | L+C | DD | 3M | **94.95±1.91** | **95.19±2.57** |
| TransFuser [42] | L+C | DD | 228K | 54.52±4.29 | 78.41±3.75 |
| VAD [49] | C | DD | 228K | 64.30 | 87.30 |
| NEAT [45] | C | DD | 130K | 58.70±4.11 | 77.32±4.91 |
| Roach [46] | C | DD | - | 65.26±3.63 | 88.24±5.16 |
| WOR [62] | C | DD | 1M | 64.79±5.53 | 87.47±4.68 |
| LBC [48] | C | DD | 157K | 30.97±4.17 | 55.01±5.14 |
| CILRS [47] | C | DD | 720K | 7.47±2.51 | 13.40±1.09 |
| VLM + GPT-4 | C | KD | 11K | 81.31±2.37 | 94.22±3.18 |
| **LeapAD (w/o Town05)** | C | KD | 11K | 75.73±1.36 | 92.10±1.44 |
| **LeapAD** | C | KD | 11K | **83.11±0.28** | **94.98±0.54** |

outlined in Section 3.3. The memory bank in our approach is configured to a default size of 9.0K, including samples collected from various towns (01-04, 06, 07, and 10). It is worth noting that samples are obtained in a closed-loop environment without human involvement. Each sample consists of the scene descriptions $D$ depicted in Section 3.2, reasoning $R$, and decisions $S$. The reasoning $R$ explains the process of decision-making based on the scene descriptions $D$ and chain of thought.

### 4.2 Implementation Details

We employ Qwen-VL-7B [28] as the VLM for scene understanding, GPT-4 as the Analytic Process for rational and logic thinking, and Qwen1.5-1.8B [60] as Heuristic Process for automatic and quick thinking in LeapAD. We use the OpenAI embedding model as the text encoder $T_e$ to extract text embedding. To fully excite our VLM's capabilities in autonomous driving, we perform SFT with the instruction-following data discussed in Section 4.1. We utilize the AdamW optimizer [61] with $\beta_1 = 0.9$ and $\beta_2 = 0.95$, coupled with a cosine decay of the learning rate, initially set to $1e^{-5}$. The batch size is set to 16, and the model is trained for 5 epochs on 8 A100 GPUs, requiring about 26 hours. The input image resolution is set at $448 \times 448$ pixels. For Heuristic Process, we conduct SFT on Qwen1.5-1.8B for 5 epochs using samples stored in the memory bank, taking about 6 hours. The training hype-parameters are consistent with the training procedure of VLM. The detailed fine-tuning process is shown in Figure 3. The dual-process decision module outputs meta-actions (e.g., "AC", "DC", "IDLE", "STOP") at a frequency of 2 HZ, which are further refined to control signals, as detailed in the Appendix A. Please refer to Appendix D for more details about the reflection mechanism and Appendix B for the performance of VLM on both simulated and real datasets.

### 4.3 Evaluation in Closed-Loop Driving

We conduct closed-loop experiments in CARLA, a popular and realistic open-source simulator, to evaluate the performance of our LeapAD. To validate the effectiveness, we conduct a comprehensive assessment in a closed-loop driving scenario on the Town05 benchmark. Our evaluation metrics include Driving Score (DS), Route Completion (RC), and Infraction Score (IS). RC signifies the proportion of the route successfully navigated by the agent, while IS indicates penalties incurred from accidents. By multiplying RC by IS, we obtain the final metric DS for evaluating our method's driving performance on a route. Table 1 compares our method with competitive methods on the Town05 Short benchmark. Specifically, we provide three different configurations to evaluate our methods comprehensively: *i)* VLM + GPT-4 represents directly using non-fine-tuned GPT-4 as the decision module along with the VLM (Qwen-VL [28]); *ii)* LeapAD (w/o Town05) represents the dual-process system with the memory bank of 9K samples accumulated from various towns (01-04, 06, 07, and 10) except Town05; *iii)* LeapAD denotes the dual-process system with the memory bank of 18K samples accumulated from various towns (01-07, and 10) and 0.1K reflection data in Town05.

As shown in Table 1, our LeapAD outperforms all other methods that rely solely on camera sensor input. Besides, our method surpasses TransFuser [42], which additionally utilizes LiDAR sensor inputs. It is worth noting that in all experiments, we only used a total of 11K data to fine-tune the

Table 2: Generalization of accumulated knowledge in the memory bank. We evaluate the performance of our LeapAD across various towns (Town05, Town01, and Town04) by leveraging memory banks accumulated from diverse sources. $L_{avg}$ denotes the average lengths of routes within each town.

| Test town | $L_{avg}$(m) | Memory (Town01-04, 06) | | Memory (Town05) | | DS ↑ | RC ↑ | IS ↑ |
|---|---|---|---|---|---|---|---|---|
| | | Few-shot | SFT | Few-shot | SFT | | | |
| Town05 | 70.1 | | ✓ | | | 66.40 | 90.40 | 73.81 |
| | | ✓ | ✓ | | | 75.73 | 92.10 | 82.66 |
| | | | | | ✓ | 69.90 | 91.79 | 76.64 |
| | | | | ✓ | ✓ | 78.07 | 91.69 | 85.89 |
| | | ✓ | ✓ | ✓ | ✓ | 83.11 | 94.98 | 87.78 |
| Town01 | 129.1 | | | ✓ | ✓ | 68.68 | 100.0 | 68.68 |
| Town04 | 119.3 | | | ✓ | ✓ | 95.08 | 97.96 | 96.56 |

VLM, while all other methods employ tens to hundreds of times more data. Moreover, our dual-process decision module does not involve human annotations, demonstrating the labeling efficiency of LeapAD. For instance, although InterFuser [41] achieves higher performance than our LeapAD, it relies on 3 million camera and LiDAR annotations, approximately 272 times more than our method. As shown in the results, we observe that even without prior driving experience in Town05, LeapAD can surpass other camera-input methods. However, there remains a gap compared to VLM + GPT-4, which integrates the scene understanding module and the Analytic Process, achieving a DS of 81.31. This demonstrates that GPT-4's understanding of world knowledge and common sense aids in performing driving-specific tasks. But using GPT-4 directly as the decision-making module is both time-consuming and expensive, making it impractical for deployment in vehicles. On the other hand, with the continuous accumulation and adaptation of experience in the test town, our LeapAD surpasses VLM + GPT-4, while only using a 1.8B model in Heuristic Process. By leveraging an enriched memory bank, it achieves a final DS of 83.11. This result fully validates the effectiveness of our dual-process decision-making module. Moreover, we provide the evaluation results on the Town05 Long benchmark and visualizations in Appendix E and F.

### 4.4 Ablation Study

We conduct extensive ablation studies about the number of few shots, size of the memory bank, reflection mechanism, and accumulated experience in a closed-loop driving setup to demonstrate the generalization and continuous learning capabilities of our LeapAD.

**Ablation on the number of few-shot.** We highlight the importance of few-shot prompting in using accumulated experience to guide current decision-making. The experiments are tested on the Town05 Short benchmark with the memory bank consisting of 9K samples automatically collected by Analytic Process in Town05. The results are presented in Figure 4, showing that our approach surpasses several methods (e.g., CLIRS [63], LBC [48] in Table 1) even with a zero-shot setting. Furthermore, there is a notable performance improvement when moving from zero-shot to one-shot scenarios. Moreover, there is a consistent increase in closed-loop experiment results as the number of shots is increased to three, which experimentally demonstrates the value of the experience in the memory bank and the effectiveness of the few-shot strategy.

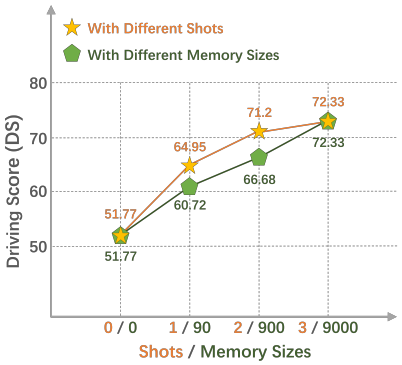



Figure 4: The illustration for ablation studies of few-shot and memory size. See Appendix E for the detailed data.

**The impact of memory sizes.** The memory bank contains accumulated experiences crucial for improving the performance of our approach. Therefore, we conduct additional ablation studies to explore the impact of the size of the memory bank, which is equipped with the few-shot strategy (defaulting to 3-shot). We conduct closed-loop evaluations with memory banks of different sizes: *i)* base memory bank with 9K samples. *ii)* compressed memory banks with 900 and 90 samples evenly sampled from the base memory bank. *iii)* no memory bank in system. The quantitative results are presented in Figure 4 and illustrate a gradual performance increase as the memory size grows. This

further demonstrates the continuous learning capability of our proposed LeapAD, indicating that our model's performance can improve with accumulated experience.

**Effectiveness of the reflection mechanism.** Reflection plays a crucial role in the continuous learning capability of our proposed LeapAD. It includes reflecting on mistakes and incorporating correct experiences into the memory bank, fostering self-improvement through proactive summarization and accumulation of experiences in unfamiliar scenes, notably corner cases. To enhance experimental efficiency, we employ a configuration of three shots and a memory bank with 900 stored memories in the Town05 environment. Reflective experiments are conducted by selecting routes with scores below 50. incorporated, and the post-reflection experiences were added to the memory bank. Figure 5 indicates that reflection significantly enhances out method's performance, albeit with some instances where individual sequence scores temporarily decrease, potentially due to the inherent randomness of the simulation environment and the limitations of the VLM model. Addressing these limitations will be the focus of our future work.

**Generalization of accumulated knowledge.** Finally, we conduct a series of experiments to demonstrate the generalization and transferability of the experience in the memory bank. The results are illustrated in Table 2, which shows the robustness of our proposed LeapAD on different towns. As shown in Row 1 and Row 3, solely employing experience adaption (SFT) from different towns with zero-shot, our LeapAD has achieved commendable performance on the Town05 Short benchmark, surpassing many methods in Table 1. Implementing the few-shot strategy yields substantial enhancements (Row 2 & 4), further highlighting the effectiveness of the experience in the memory bank. Furthermore, we perform cross-validation to demonstrate the generalization of the accumulated experience, focusing on two setups: *i)* testing on Town05 with the memory bank collected from other towns (Row 1 & 2); *ii)* testing on the other towns (Town01, 04) with the memory bank accumulated only from Town05 (Row 6 & 7). Comparing Row 2 with Row 4, our LeapAD utilizing different memory sources achieves comparable performance in the same town (Town05). Additionally, our proposed LeapAD demonstrates good performance across different towns when using the same memory bank (Row 4, 6 & 7). In this section, several low-quality samples in the memory bank on routes with low driving scores are withdrawn when accumulating experience, and no special memory processing is performed on those experiments in Figure 4.

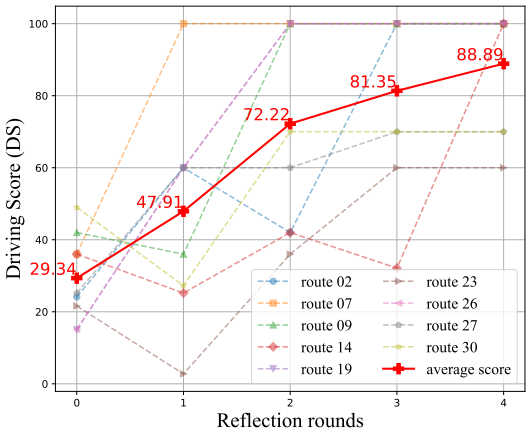



Figure 5: Effectiveness of the reflection mechanism. The $x$-axis represents the rounds of reflection, while the $y$-axis denotes the resulting driving score. The dashed line illustrates performances on different routes after multi-round reflection, and the red "average score" denotes the mean performance across all routes.

## 5 Conclusion

In this paper, we introduce LeapAD, a dual-process closed-loop autonomous driving system with continuous learning, adapting, and improving capabilities. Similar to human attention, our approach selectively prioritizes critical objects that can influence driving decisions, simplifying the scene description and reducing decision-making complexity. Furthermore, the dual-process decision-making module mimics human cognitive processes through a fast, empirical Heuristic Process and a slow, rational Analytic Process. Through reflection mechanisms and a transferable memory bank, LeapAD continuously improves from past experiences in a closed-loop environment, demonstrating continuous learning capabilities and strong adaptability to various driving scenarios. Moreover, LeapAD can be seamlessly integrated with the mainstream cloud-edge architectures employed in intelligent vehicles. The Heuristic Process operates at the edge, enabling instant decision-making within the vehicle, while the Analytic Process handles more complex scenarios in the cloud.

## 6 Limitations and Broader Impacts

Currently, LeapAD relies solely on single-frame camera inputs, without any temporal input. Another bottleneck of our approach is the VLM's inability to participate in the reflection mechanism, which hinders further system improvements. Additionally, there is a notable gap between predefined agent behaviors in the CARLA benchmark and those in real-world scenarios, underscoring the need for a high-fidelity world simulator. For impacts, autonomous driving systems gather extensive data on driving behaviors, routes, and passenger movements, raising concerns about data privacy and legal implications. Nonetheless, advancements in technology and regulatory frameworks can help address these issues, paving the way for safer, more efficient, and accessible driving systems.

## 7 Acknowledgments

This research was supported by the NSFC 62088101 Autonomous Intelligent Unmanned Systems. Additionally, it received support from the Shanghai Artificial Intelligence Laboratory, the National Key R&D Program of China (Grant No. 2022ZD0160104), and the Science and Technology Commission of Shanghai Municipality (Grant No. 22DZ1100102).

## References

[1] S. Thrun, M. Montemerlo, H. Dahlkamp, D. Stavens, A. Aron, J. Diebel, P. Fong, J. Gale, M. Halpenny, G. Hoffmann, *et al.*, "Stanley: The robot that won the darpa grand challenge," *Journal of field Robotics*, vol. 23, no. 9, pp. 661–692, 2006.

[2] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11784–11793, 2021.

[3] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*, pp. 1–18, Springer, 2022.

[4] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*, pp. 2774–2781, IEEE, 2023.

[5] Wayve, "Lingo-1: Exploring natural language for autonomous driving," 2023.

[6] Y. Ma, Y. Cao, J. Sun, M. Pavone, and C. Xiao, "Dolphins: Multimodal language model for driving," *arXiv preprint arXiv:2312.00438*, 2023.

[7] X. Li, Y. Bai, P. Cai, L. Wen, D. Fu, B. Zhang, X. Yang, X. Cai, T. Ma, J. Guo, *et al.*, "Towards knowledge-driven autonomous driving," *arXiv preprint arXiv:2312.04316*, 2023.

[8] Y. Zhou, L. Huang, Q. Bu, J. Zeng, T. Li, H. Qiu, H. Zhu, M. Guo, Y. Qiao, and H. Li, "Embodied understanding of driving scenarios," *arXiv preprint arXiv:2403.04593*, 2024.

[9] H. Shao, Y. Hu, L. Wang, S. L. Waslander, Y. Liu, and H. Li, "Lmdrive: Closed-loop end-to-end driving with large language models," *arXiv preprint arXiv:2312.07488*, 2023.

[10] J. Mao, J. Ye, Y. Qian, M. Pavone, and Y. Wang, "A language agent for autonomous driving," *arXiv preprint arXiv:2311.10813*, 2023.

[11] J. Yuan, S. Sun, D. Omeiza, B. Zhao, P. Newman, L. Kunze, and M. Gadd, "Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model," *arXiv preprint arXiv:2402.10828*, 2024.

[12] X. Tian, J. Gu, B. Li, Y. Liu, C. Hu, Y. Wang, K. Zhan, P. Jia, X. Lang, and H. Zhao, "Drivevlm: The convergence of autonomous driving and large vision-language models," *arXiv preprint arXiv:2402.12289*, 2024.

[13] Z. Li, Z. Yu, S. Lan, J. Li, J. Kautz, T. Lu, and J. M. Alvarez, "Is ego status all you need for open-loop end-to-end autonomous driving?," *arXiv preprint arXiv:2312.03031*, 2023.

[14] D. Kahneman, "Fast and slow thinking," *Allen Lane and Penguin Books, New York*, 2011.

[15] J. S. B. Evans and K. E. Stanovich, "Dual-process theories of higher cognition: Advancing the debate," *Perspectives on psychological science*, vol. 8, no. 3, pp. 223–241, 2013.

[16] P. C. Wason and J. S. B. Evans, "Dual processes in reasoning?," *Cognition*, vol. 3, no. 2, pp. 141–154, 1974.

[17] S. Epstein, "Cognitive-experiential self-theory of personality," *Comprehensive handbook of psychology*, vol. 5, pp. 159–184, 2003.

[18] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[19] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[20] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, "Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality," March 2023.

[21] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23716–23736, 2022.

[22] X. Chen, X. Wang, S. Changpinyo, A. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, A. Grycner, B. Mustafa, L. Beyer, *et al.*, "Pali: A jointly-scaled multilingual language-image model," *arXiv preprint arXiv:2209.06794*, 2022.

[23] B. Li, Y. Zhang, L. Chen, J. Wang, F. Pu, J. Yang, C. Li, and Z. Liu, "Mimic-it: Multi-modal in-context instruction tuning," *arXiv preprint arXiv:2306.05425*, 2023.

[24] K. Li, Y. He, Y. Wang, Y. Li, W. Wang, P. Luo, Y. Wang, L. Wang, and Y. Qiao, "Videochat: Chat-centric video understanding," *arXiv preprint arXiv:2305.06355*, 2023.

[25] Z. Li, B. Yang, Q. Liu, Z. Ma, S. Zhang, J. Yang, Y. Sun, Y. Liu, and X. Bai, "Monkey: Image resolution and text label are important things for large multi-modal models," *arXiv preprint arXiv:2311.06607*, 2023.

[26] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[27] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, *et al.*, "Cogvlm: Visual expert for pretrained language models," *arXiv preprint arXiv:2311.03079*, 2023.

[28] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A frontier large vision-language model with versatile abilities," *arXiv preprint arXiv:2308.12966*, 2023.

[29] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, Z. Muyan, Q. Zhang, X. Zhu, L. Lu, *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," *arXiv preprint arXiv:2312.14238*, 2023.

[30] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[31] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[32] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.

[33] D. Fu, X. Li, L. Wen, M. Dou, P. Cai, B. Shi, and Y. Qiao, "Drive like a human: Rethinking autonomous driving with large language models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 910–919, 2024.

[34] L. Wen, D. Fu, X. Li, X. Cai, T. MA, P. Cai, M. Dou, B. Shi, L. He, and Y. Qiao, "Dilu: A knowledge-driven approach to autonomous driving with large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[35] L. Wen, X. Yang, D. Fu, X. Wang, P. Cai, X. Li, T. MA, Y. Li, L. XU, D. Shang, Z. Zhu, S. Sun, Y. BAI, X. Cai, M. Dou, S. Hu, B. Shi, and Y. Qiao, "On the road with GPT-4v(ision): Explorations of utilizing visual-language model as autonomous driving agent," in *ICLR 2024 Workshop on Large Language Model (LLM) Agents*, 2024.

[36] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, P. Luo, A. Geiger, and H. Li, "Drivelm: Driving with graph visual question answering," *arXiv preprint arXiv:2312.14150*, 2023.

[37] A.-M. Marcu, L. Chen, J. Hünermann, A. Karnsund, B. Hanotte, P. Chidananda, S. Nair, V. Badrinarayanan, A. Kendall, J. Shotton, E. Arani, and O. Sinavski, "Lingoqa: Video question answering for autonomous driving," *arXiv preprint arXiv:2312.14115*, 2023.

[38] M. Nie, R. Peng, C. Wang, X. Cai, J. Han, H. Xu, and L. Zhang, "Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving," *arXiv preprint arXiv:2312.03661*, 2023.

[39] W. Wang, J. Xie, C. Hu, H. Zou, J. Fan, W. Tong, Y. Wen, S. Wu, H. Deng, Z. Li, *et al.*, "Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving," *arXiv preprint arXiv:2312.09245*, 2023.

[40] X. Jia, P. Wu, L. Chen, J. Xie, C. He, J. Yan, and H. Li, "Think twice before driving: Towards scalable decoders for end-to-end autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21983–21994, 2023.

[41] H. Shao, L. Wang, R. Chen, H. Li, and Y. Liu, "Safety-enhanced autonomous driving using interpretable sensor fusion transformer," in *Conference on Robot Learning*, pp. 726–737, PMLR, 2023.

[42] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[43] P. Wu, X. Jia, L. Chen, J. Yan, H. Li, and Y. Qiao, "Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline," *Advances in Neural Information Processing Systems*, vol. 35, pp. 6119–6132, 2022.

[44] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17853–17862, 2023.

[45] K. Chitta, A. Prakash, and A. Geiger, "Neat: Neural attention fields for end-to-end autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15793–15803, 2021.

[46] Z. Zhang, A. Liniger, D. Dai, F. Yu, and L. Van Gool, "End-to-end urban driving by imitating a reinforcement learning coach," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15222–15232, 2021.

[47] F. Codevilla, E. Santana, A. M. López, and A. Gaidon, "Exploring the limitations of behavior cloning for autonomous driving," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9329–9338, 2019.

[48] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," in *Conference on Robot Learning*, pp. 66–75, PMLR, 2020.

[49] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Vad: Vectorized scene representation for efficient autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8350, 2023.

[50] Y. Peng, J. Han, Z. Zhang, L. Fan, T. Liu, S. Qi, X. Feng, Y. Ma, Y. Wang, and S.-C. Zhu, "The tong test: Evaluating artificial general intelligence through dynamic embodied physical and social interactions," *Engineering*, 2023.

[51] S. Gildert and G. Rose, "Building and testing a general intelligence embodied in a humanoid robot," *arXiv preprint arXiv:2307.16770*, 2023.

[52] Z. Liu, H. Jiang, H. Tan, and F. Zhao, "An overview of the latest progress and core challenge of autonomous vehicle technologies," in *MATEC Web of Conferences*, vol. 308, p. 06002, EDP Sciences, 2020.

[53] F. Dou, J. Ye, G. Yuan, Q. Lu, W. Niu, H. Sun, L. Guan, G. Lu, G. Mai, N. Liu, *et al.*, "Towards artificial general intelligence (agi) in the internet of things (iot): Opportunities and challenges," *arXiv preprint arXiv:2309.07438*, 2023.

[54] B. Zhang, J. Zhu, and H. Su, "Toward the third generation artificial intelligence," *Science China Information Sciences*, vol. 66, no. 2, p. 121101, 2023.

[55] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, *et al.*, "The rise and potential of large language model based agents: A survey," *arXiv preprint arXiv:2309.07864*, 2023.

[56] Y. Liu, F. Wu, Z. Liu, K. Wang, F. Wang, and X. Qu, "Can language models be used for real-world urban-delivery route optimization?," *The Innovation*, vol. 4, no. 6, 2023.

[57] C. Cui, Y. Ma, X. Cao, W. Ye, Y. Zhou, K. Liang, J. Chen, J. Lu, Z. Yang, K.-D. 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.

[58] E. Sachdeva, N. Agarwal, S. Chundi, S. Roelofs, J. Li, M. Kochenderfer, C. Choi, and B. Dariush, "Rank2tell: A multimodal driving dataset for joint importance ranking and reasoning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 7513–7522, 2024.

[59] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*, pp. 1–16, PMLR, 2017.

[60] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[61] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[62] D. Chen, V. Koltun, and P. Krähenbühl, "Learning to drive from a world on rails," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15590–15599, 2021.

[63] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[64] R. C. Coulter *et al.*, *Implementation of the pure pursuit path tracking algorithm*. Carnegie Mellon University, The Robotics Institute, 1992.

[65] T. Wang, E. Xie, R. Chu, Z. Li, and P. Luo, "Drivecot: Integrating chain-of-thought reasoning with end-to-end driving," *arXiv preprint arXiv:2403.16996*, 2024.

# Appendix

## A Low-level Control

As mentioned, our dual-process decision-making module outputs meta-actions (e.g., "AC", "DC", "IDLE", "STOP"), which are further refined to control signals such as steering, acceleration and brake. For the sake of simplicity, we define "AC" to mean acceleration of 1m/s, and "DC" to mean deceleration of -1m/s.

### A.1 Planned Waypoints

The default route waypoints provided by CARLA are sparse, with distances between consecutive waypoints reaching up to several dozen meters. This makes it difficult for our lower-level controller to decompose meta-actions into control signals. To address this issue, we leverage the high-definition map to densify these sparse waypoints into 1-meter interval path-points, which form the reference path for the ego vehicle. Subsequently, our controller employs the Pure Pursuit algorithm [64] to track the reference path, ensuring that the ego vehicle remains on the correct road. The selection of the target path-point for tracking is adaptive, depending on the vehicle's speed, with the controller choosing one of the third to seventh path points ahead.

It is worth noting that high-definition maps are not a necessary requirement for our approach. Alternative methods, such as those proposed in DriveCot [65] or TransFuser [42], which utilize separate neural networks to predict the future reference path based on camera images with sparse navigation information, are also compatible with controller design, without affecting our core methodology.

### A.2 PID Controller

Given the target path-point and the target vehicle speed, ego vehicle's control method employs two independent PID controllers, building upon previous works [45, 42, 43, 40, 65]. Specifically, the control system contains two separate PID controllers: one longitudinal for throttle and brake and one lateral for steering. The longitudinal PID controller, tuned with gains $K_P = 5.0$, $K_I = 0.5$, and $K_D = 1.0$, takes the current speed and desired target speed as inputs and utilizes a 40-frame buffer to compute the throttle and brake values. Meanwhile, the lateral PID controller, tuned with gains $K_P = 1.0$, $K_I = 0.5$, and $K_D = 0.2$, receives the angle difference between the ego vehicle's heading and the vector pointing to the chosen future waypoint as input, and employs a 20-frame buffer to calculate the steering value.

## B Dataset Annotations

In this section, we present examples of datasets utilized to train the VLM, including the self-collected CARLA dataset (Figure 6), the DriveLM dataset (Figure 7), and the Rank2Tell dataset (Figure 8). The data annotations principally categorize the critical objects based on four attributes (semantic, spatial, motion, and risk ranking), as outlined in Figure 9 and Section 3.2. It is worth noting that not every dataset comprehensively covers questions and answers pertaining to these attributes. For instance, the DriveLM dataset lacks the selection of important objects; the Rank2Tell dataset merges semantic and spatial information into a single overview.

To conduct closed-loop experiments, we specifically define questions about the semantic and spatial details of key objects in the CARLA dataset and design several automatic annotation rules. Specifically, the critical objects are defined as those situated in the front, left, and right fields of view: *i)* Vehicles and cyclists positioned within 20 meters of the ego car or less than 60 meters in the ego lane. *ii)* Pedestrians within 40 meters of the ego car. *iii)* Traffic lights that control the vehicle's travel direction and stop signs.

In addition to presenting the datasets, we further elaborate on the description generated by the VLM. The examples described in Figure 6, 7, 8 demonstrate VLM's proficiency in accurately identifying key objects that influence driving decisions and effectively describing the related attributes. We observe that VLM not only excels in simulated environments but also demonstrates robust performance

Figure 6: Annotation format and example of scene description generated by VLM on a self-collected CARLA simulation dataset

Figure 7: Annotation format and example of scene description generated by VLM on DriveLM dataset.

Figure 8: Annotation format and example of scene description generated by VLM on Rank2Tell dataset. It demonstrates the robust performance of our VLM in real-world scenarios

in real-world scenarios. The versatility of our scene understanding module enables the proposed LeapAD with the promise of practical applicability across real-world scenarios.

## C Prompt Details

We outline the specifics of the system prompt (Figure 10) utilized by our Analytic Process during the accumulation of experience within a closed-loop environment. The prompt consists of task definitions, meta-actions, adherence to traffic rules, and the desired output format. Furthermore, the figure illustrates the system prompt (Figure 11) utilized during the reflection procedure, which contains fundamental descriptions from the previous prompt alongside criteria for identifying potential errors within historical frames. We also detail the prompts (Figure 12) of the VLM utilized for identifying the critical objects in the traffic scenes.

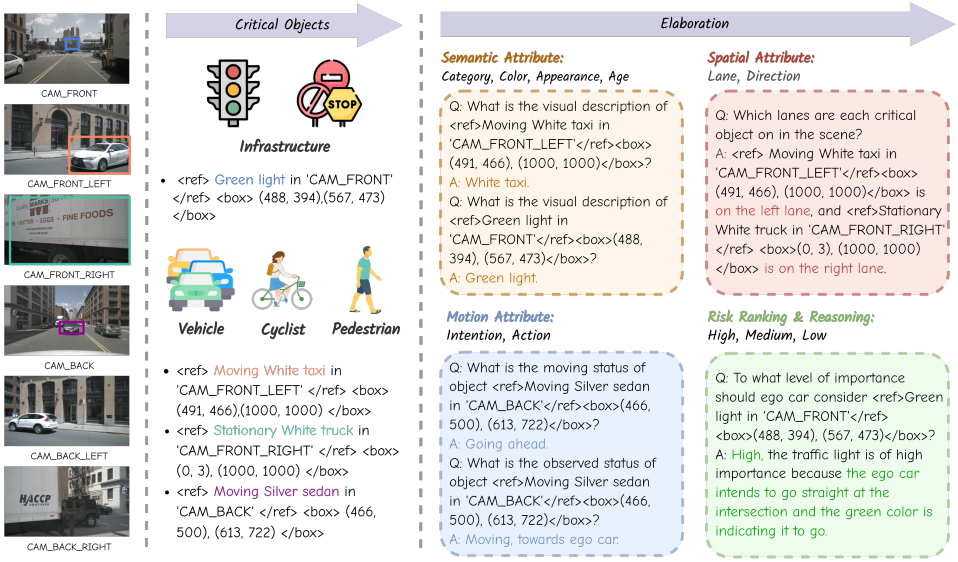



Figure 9: Detailed data format for the descriptions of critical objects.

```
SYSTEM_PROMPT = '''
You are a large multi-modal model trained by OpenAI. Now you act as a mature driving assistant, who can give accurate and correct advice for human driver in complex urban driving scenarios. You'll receive some scene description from the view of onboard camera. You'll need to make driving inferences and decisions based on the information. At each decision frame, you receive information about the current scene and a collection of actions. You will perform reasoning based on the the description of front-view image. Eventually you will select the appropriate action output from the action set.
Make sure that all of your reasoning is output in the `## Reasoning` section, and in the `## Decision` section you should only output the name of the action, e.g. `AC`, `IDLE` etc.

## Available Actions
 - AC: increasing your speed.
 - DC: slow down your speed.
 - IDLE: maintain your current speed.
 - STOP: stop, your current speed shold be zero.

You must obey these important rules below:
- You should pay more attention to the object on the ego lane. Vehicles that are not on the ego lane that are moving towards the ego car bring less potential risk. If they are always moving on the left or right lane, you don't need to slow down. You need to notice if they change to the ego lane and come to a close, then you need to slow down or stop to keep your distance and avoid the potential collision.
- When your steering value is negative, you should pay attention to objects in the left lane; when your steering value is positive, you should pay attention to objects in the right lane.
- You should always keep your distance from the objects around you, especially the object on the ego lane. You should 'DC' or even 'STOP' when an object is very close on the ego lane
- If you see a red light on the ego lane from the view, you should 'DC' or 'STOP' immediately. You should stop just before the white line (perpendicular to the direction you are running) in front of you and do not cross it.
- If there is any object in the view that is less than 25m away from the vehicle, you should slow down ('DC'), and if the distance is less than 15 meters, you should 'STOP' immediately to avoid collision. If the distances of all objects are large than 25m, there are no red traffic lights and stop signs in front of you, and you are at low speeds (e.g. lower than 1m/s), you should select "AC" insteand of "IDLE".
- If there is any stop sign that is less than 10m away, you should 'DC' and if it is less than 5m away from the vehicle, you should 'STOP' immediately.

Your answer should follow this format:

## Reasoning
reasoning based on the descrioption of the front-view.

## Decison
one of the actions in the action set.(SHOULD BE exactly same and no other words!)

You must obey the rules above. \n
'''
```

Figure 10: System prompt for Analytic Process

```
REFLECTION_SYSTEM_PROMPT = """
You are a part of an automotive safety analysis team. Your task is to reevaluate reasoning and decisions based on current vehicle accident information and descriptions of historical frames. You should identify potential inference errors in historical frames and propose revised inferences and decisions based on this reevaluation. Note that you should select one historical frame that has the best chance of avoiding the current accident to suggest your modifications about the reasoning and decision.
Make sure that all of your reasoning is output in the `## Reasoning` section, and in the `## Decision` section you should only output the name of the action, e.g. `AC`, `IDLE` etc.

## Available Actions
- AC: increasing your speed.
- DC: slow down your speed.
- IDLE: maintain your current speed.
- STOP: stop, your current speed shold be zero.

You must obey these important rules below:
- You should pay more attention to the object on the ego lane. Vehicles that are not on the ego lane that are moving towards the ego car bring less potential risk. If they are always moving on the left or right lane, you don't need to slow down. You need to notice if they change to the ego lane and come to a close, then you need to slow down or stop to keep your distance and avoid the potential collision.
- When your steering value is negative, you should pay attention to objects in the left lane; when your steering value is positive, you should pay attention to objects in the right lane.
- You should always keep your distance from the objects around you, especially the object on the ego lane. You should 'DC' or even 'STOP' when an object is very close on the ego lane
- If you see a red light on the ego lane from the view, you should 'DC' or 'STOP' immediately. You should stop just before the white line (perpendicular to the direction you are running) in front of you and do not cross it.
- If there is any object in the view that is less than 25m away from the vehicle, you should slow down ('DC'), and if the distance is less than 15 meters, you should 'STOP' immediately to avoid collision. If the distances of all objects are large than 25m, there are no red traffic lights and stop signs in front of you, and you are at low speeds (e.g. lower than 1m/s), you should select "AC" instead of "IDLE".
- If there is any stop sign that is less than 10m away, you should 'DC' and if it is less than 5m away from the vehicle, you should 'STOP' immediately.

Your answer should follow this format:

## Time
time of the selected historical frame \n

## Reasoning
reasoning about the decision based on the descrioption of the selected frame. Pay attention to modifying pronouns, such as using "this frame" or "this moment" instead of using "the historical frame" or "at time of".

## Decision
one of the actions in the action set.(SHOULD BE exactly same and no other words!)

You must obey the rules above. \n
"""
```

Figure 11: System prompt in the reflection mechanism

```
VLM_SYSTEM_PROMPT_TEMPLATE=
  """Current scene: {}, <img>{}</img>. You are driving a car in an urban street and this image indicates the scene that you see. Please describe in detail the key objects in the scene that may affect your driving. The object information should contain the object's category, the object's location relative to the ego car or the ego lane, the object's motion state, and the object's bounding box, as well as the approximate distance from the ego car. Especially, if there are any traffic lights, the information should contain the traffic light's color and bounding box."""
```

Figure 12: VLM system prompt

Table 3: Comparison of our LeapAD with competitive methods in Town05 Long benchmark. "DD" and "KD" denote data-driven and knowledge-driven, respectively.

| Method | Modality | Type | Annotations | DS (%) ↑ | RC (%) ↑ | IS (%) ↑ |
|---|---|---|---|---|---|---|
| DriveMLM [39] | L+C | DD | 2M | 76.1 | 98.1 | 78.0 |
| ThinkTwice [40] | L+C | DD | 2M | 70.9 | 95.5 | 75.0 |
| InterFuser [41] | L+C | DD | 3M | 68.3 | 95.0 | 72.0 |
| TransFuser [42] | L+C | DD | 228K | 31.0 | 47.5 | 77.0 |
| VAD [49] | C | DD | 228K | 30.3 | 75.2 | - |
| TCP [43] | C | DD | 420K | 57.2 | 80.4 | 73.0 |
| NEAT [45] | C | DD | 130K | 37.7 | 62.1 | 61.0 |
| Roach [46] | C | DD | - | 43.6 | 80.4 | 54.4 |
| WOR [62] | C | DD | 1M | 44.8 | 82.4 | 54.0 |
| LBC [48] | C | DD | 157K | 7.1 | 32.1 | 22.1 |
| CILRS [47] | C | DD | 720K | 3.7 | 7.2 | 51.4 |
| LeapAD (ours) | C | KD | 11K | 51.7 | 100 | 51.7 |

Table 4: Ablation on the number of few shots.

| Shots | DS (%) ↑ | RC (%) ↑ | IS (%) ↑ |
|---|---|---|---|
| 0 | 51.77 | 99.72 | 52.45 |
| 1 | 64.95 | 100.0 | 64.95 |
| 2 | 71.20 | 94.86 | 73.60 |
| 3 | **72.33** | **100.0** | **72.33** |

Table 5: Affect of the memory size for few-shot learning.

| Mem Sizes | DS (%) ↑ | RC (%) ↑ | IS (%) ↑ |
|---|---|---|---|
| 0 | 51.77 | 99.72 | 52.45 |
| 90 | 60.72 | 98.30 | 62.30 |
| 900 | 66.68 | 97.93 | 68.10 |
| 9000 | **72.33** | **100.0** | **72.33** |

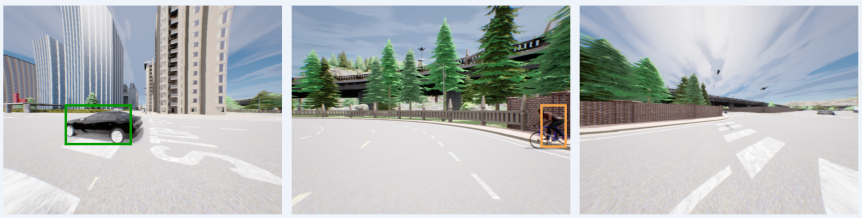
## D Reflection mechanism

As mentioned in Section 3.3, we employ the Analytic Process to reflect on traffic accidents, facilitating the system's capabilities to continuously improve. Specifically, in the reflection mechanism, we maintain a memory queue $\boldsymbol{Q} = \{(D_i, R_i, S_i)\}_{i=0}^{9}$, storing the previous 10 samples from Heuristic Process at a frequency of 1 Hz. When an accident occurs, samples in $\boldsymbol{Q}$ are fed to Analytic Process for reflection. This prompts the system to identify keyframes that caused the accident and provide correct reasoning and decisions, which are then added to the memory bank. The detailed prompt used in the reflection procedure is depicted in Figure 11. In addition, Figure 16 depicts a specific reflection procedure.

## E Other experiments

Moreover, we provide the evaluation results on the Town05 Long benchmark in Table 3. The results show that our LeapAD can still achieve competitive results compared to those methods that take images as input, despite using annotated data that is much smaller. Our study delves into the nuanced effects of few-shot learning and memory size through rigorous ablation analyses, as detailed in Tables 4 and 5. The baseline performance of the SFT model without memory is presented in the first row, revealing a clear positive correlation with increased shots and expanded memory capacity in closed-loop experiments. Table 4 substantiates the efficacy of the few-shot heuristic process and highlights the success of our method in identifying similar memories. Table 5 demonstrates the incremental performance gains as memory accumulates, exemplifying the principle of continuous learning.

Figure 13: Case study at the intersection.



## Scene Description

There are 2 important objects in the scene. The information of these key objects are: <ref> In 'CAM_FRONT', vehicle is located approximately 10.574 m to the front right of the ego car is driving in the same direction </ref><box> (881, 465), (913, 687) </box><ref> In 'CAM_LEFT', vehicle is located approximately 7.673 m to the back left of the ego car is driving in the opposite direction </ref><box> (197, 476), (509, 671) </box>.

## Ego State

Your current speed is 6.238167388988819 m/s. Your current steering value is -0.18778696061027783.

## Reasoning

There is a vehicle on the front right of the ego car at a distance of approximately 10.574 meters, which is close enough to warrant attention as it is within the same lane or close to it. The relative coordinates show that this vehicle is mostly in the right lane, not directly in the ego lane. Given the current steering value is negative, we should also be aware of objects on the left side, but since the vehicle in the left lane is moving in the opposite direction and no other details require immediate attention, it does not influence the decision-making in this scenario. Since the vehicle in the right lane is relatively close and any sudden lane change into the ego lane could present a risk, it would be prudent to slow down to maintain a safe following distance and be ready to react if necessary. There are no red traffic lights or stop signs mentioned, so those rules do not apply.

## Decision

DC

Figure 14: A case where a bicycle suddenly appears.

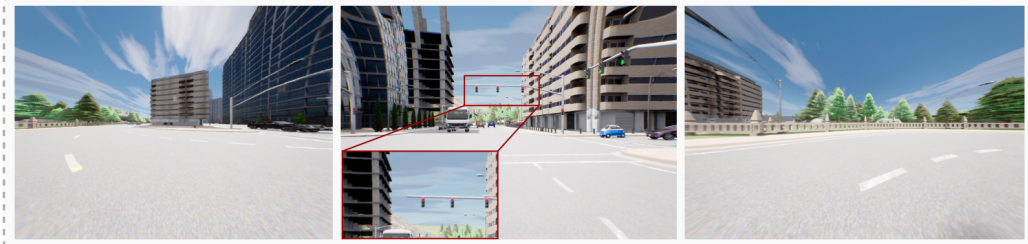
## F Visualization Cases

**Representative cases.** We also provide several examples to show the qualitative performance of our proposed LeapAD, as illustrated in Figure 13, 14. These cases present the zero-shot results of our proposed system: The VLMs take the images of the current scene and generate the scene descriptions, which are fed into the decision module (Heuristic Process) along with the ego state (i.e., speed and steering values) for reasoning and decision-making. As shown in Figure 13, VLM correctly perceives the important traffic participants at the intersection, such as the red lights and passing vehicles, prompting the decision module to make a deceleration decision. Figure 14 showcases a challenging scenario where a cyclists suddenly appears from the curb while the vehicle is moving. The system adeptly detects the risk and promptly responds with a timely deceleration.

**Cases of few-shot prompting.** Figure 15 presents a case to show the results using the few-shot strategy. As mentioned in Section 3.4, several memory samples are queried based on similarities of the scene descriptions' embedding between the current scene and those in the memory bank. From this case, we can see that the queried samples are highly related to the current scene. And our system LeapAD observes there is a red traffic light and several vehicles in the front. According to the previous experience and current ego status, the system suggests that it should obey the red traffic light and prepare to stop.

**Cases of reflection mechanism.** We also provide an example to illustrate the reflection mechanism in Figure 16. When an incident occurs, the historical descriptions, reasoning, and decisions are fed into the Analytic Process to analyze the sequence of events meticulously, identify potential errors, and provide the reflected reasoning and decision. As the case shows, after the collision occurs at the current frame (frame 0), the Analytic Process found there exist reasoning errors at the previous frame (frame -2). For instance, when a vehicle is dangerously close (6.98m) ahead of the ego car, the initial analysis by the Heuristic Process incorrectly identifies the crucial object and misinterprets its movement status, necessitating immediate attention. Furthermore, the Heuristic Process struggles to comprehend the relationship between the ego car's speed and the safe distance from the vehicle in front, resulting in poor decision-making. Conversely, during the reflection procedure, the Analytic Process accurately identifies the vehicle 6.98 meters ahead as the primary concern. It questions the prior decision to maintain an "IDLE" status, suggesting it may have been an inference error, as stopping would have been necessary to avoid closing in too rapidly on the vehicle ahead in this scenario.

**Failure Cases.** We have included two typical failure cases in the uploaded PDF. (1) "Run a red light" as shown in Figure 17. In this scenario, the system lacks temporal information regarding the yellow light's remaining duration, making it difficult to determine whether to accelerate through or stop. When the light is yellow, the system cautiously issues a "DC" command, causing the vehicle to cross the stop line slowly. When the light turned red, CARLA interpreted this as running a red light, even though a "STOP" command was issued at this time. (2) "Collision" as shown in Figure 18. In this case, the VLM did not detect the car at the left rear edge of the field of view due to the camera's field of view limitation. Furthermore, in the CARLA setting, other vehicles will not proactively yield to the ego vehicle, leading to collisions caused by other vehicles.



## Current frame

## Scene Description
There are **4** important objects in the scene. The information of these key objects are: <ref> red traffic light is located approximately 20.845 m to the front of the ego car </ref> <box> (529, 342), (535, 370) </box> <ref> vehicle is located approximately 47.668 m to the front right of the ego car is driving in the same direction </ref> <box> (504, 490), (528, 518) </box> <ref> vehicle is located approximately 36.513 m to the front left of the ego car is away from ego to the left </ref> <box> (391, 494), (423, 523) </box> <ref> vehicle is located approximately 32.154 m to the front left of the ego car is turn left </ref> <box> (292, 427), (395, 542) </box>.

## Ego state
Your current speed is 6.744 m/s. Your current steering value is -0.0658.

## Reasoning
There is a red traffic light approximately 20.845 meters to the front of the ego car, which requires stopping the vehicle to comply with traffic regulations. Even though our current speed is relatively low (6.744 m/s), we must decelerate to come to a complete stop due to the red traffic light. Additionally, observing the traffic light confirms it is red and we must abide by it, which still warrants stopping the vehicle to some extent.

## Decision
STOP

## Queried memory

## Scene Description
There are **3** important objects in the scene. The information of these key objects are: <ref> red traffic light is located approximately 24.985 m to the front of the ego car </ref> <box> (504, 248), (514, 289) </box> <ref> vehicle is located approximately 43.446 m to the front right of the ego car is drive in the same direction </ref> <box> (503, 491), (523, 520) </box> <ref> vehicle is located approximately 38.473 m to the front left of the ego car is away from ego to the left </ref> <box> (379, 489), (431, 520) </box>.

## Ego state
Your current speed is 0.09358620085244608 m/s. Your current steering value is 0.0.

## Reasoning
There is a red traffic light in both the front view at approximately 24.985 meters ahead of the ego car. According to the scenario description and rules, we must obey the rules of the traffic lights in the view. Since we are approaching a traffic intersection, we must proceed at a low speed. Our current speed is already low (0.09358620085244608 m/s), but as we are approaching the intersection with a red light ahead, we should prepare to stop before the white line and ensure we do not enter the intersection.

## Decision
STOP

1

## Scene Description
There are **4** important objects in the scene. The information of these key objects are: <ref> red traffic light is located approximately 29.475 m to the front of the ego car </ref> <box> (504, 294), (514, 326) </box> <ref> vehicle is located approximately 30.571 m to the front right of the ego car is away from ego to the right </ref> <box> (739, 486), (866, 537) </box> <ref> vehicle is located approximately 38.217 m to the front left of the ego car is come closer from the left </ref> <box> (378, 483), (444, 527) </box> <ref> vehicle is located approximately 17.223 m to the front left of the ego car is turn right </ref> <box> (0, 478), (166, 606) </box>.

## Ego state
Your current speed is 3.822301748594917 m/s. Your current steering value is 0.0.

## Reasoning
We have multiple inputs to consider:
1. There is a red traffic light ahead, observed at approximately 29m, which takes precedence over the front view.
2. Our current speed is above the recommended low speed for approaching a traffic intersection, which is less than 3 m/s.
3. There are two vehicles in proximity, one on the front-right at about 31m which poses no immediate threat as it's moving away, and another on the front-left at about 17m which is turning right and may cross into the ego lane, posing a potential risk.
4. Considering the distance to the red light, we need to start slowing down to stop in time and avoid running the red light.
5. Our steering value is currently neutral (0.0), so we don't have to focus on one side versus the other at this moment.

Given the proximity of the red light and its precedence, together with the recommended low speed for approaching a traffic intersection, we must ignore the speed rule for this scenario and focus on obeying the traffic light and preparing to stop.

## Decision
DC

2

## Scene Description
There are **4** important objects in the scene. The information of these key objects are: <ref> red traffic light is located approximately 37.25 m to the front right of the ego car </ref> <box> (594, 393), (602, 429) </box> <ref> vehicle is located approximately 47.954 m to the front right of the ego car is driving in the same direction </ref> <box> (554, 544), (583, 574) </box> <ref> vehicle is located approximately 34.744 m to the front left of the ego car is away from ego to the left </ref> <box> (456, 556), (500, 589) </box> <ref> vehicle is located approximately 36.791 m to the front right of the ego car is come closer from the right </ref> <box> (566, 535), (617, 570) </box>.

## Ego state
Your current speed is 0.04411071693813459 m/s. Your current steering value is 0.0.

## Reasoning
From the front view, there's a red traffic light approximately 37.25 meters to the front right of the ego car, and since we are approaching a traffic intersection, we should obey the red light. There is also a vehicle about 47.95 meters to the front right driving in the same direction, a vehicle about 36.79 meters to the front right coming closer from the right, and another vehicle approximately 34.74 meters to the front left moving away. Given that the closest object is the red traffic light and it is obeying the traffic rule when the steering value is at 0.0, indicating we are in the ego lane. We need to prepare to stop for the red light.

Our current speed is extremely low at 0.044 m/s, but since the rules suggest that we must obey the red traffic light and prepare to stop when we are approaching an intersection, increasing speed is not an option. Thus, we will maintain the current speed and approach the traffic light cautiously. As the red traffic light is not imminently close (not under 25 meters), we are not required to stop immediately but should be prepared to do so.

## Decision
IDLE

3

Figure 15: Case study for few-shot strategy.

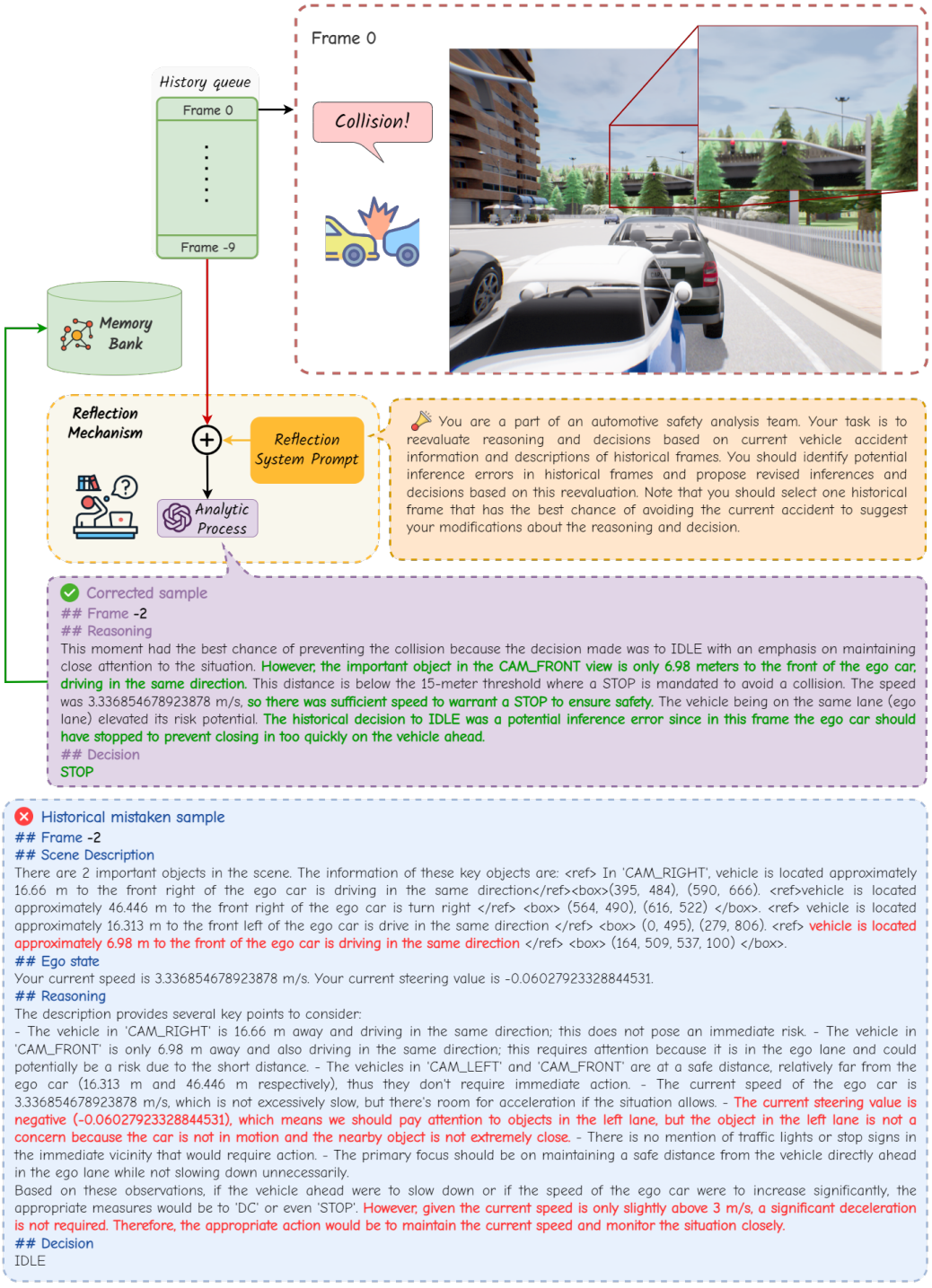
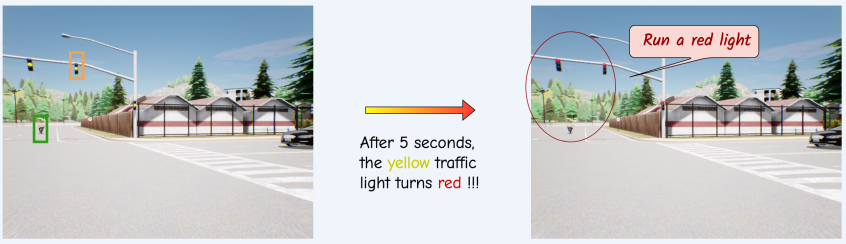



Figure 16: Case study for reflection mechanism.




## Scene Description

There are 2 important objects in the scene. The information of these key objects are: <ref> In 'CAM_FRONT', yellow traffic light is located approximately 14.015 m to the front right of the ego car </ref> <box> (320, 156), (379, 356) </box>, In 'CAM_FRONT', vehicle is located approximately 19.784 m to the front left of the ego car is away from ego to the left</ref> <box> (245, 469), (246, 543) </box>.

## Ego State

Your current speed is 0.13408580607243586 m/s. Your current steering value is 0.12698765937922699.

## Reasoning

... Additionally, the ego car's current speed is very low (0.134 m/s), which is below the threshold where time wasting at low speeds should be considered. However, the presence of a yellow traffic light approximately 14 meters to the front right indicates that action will soon be necessary at the intersection. Given the proximity of the yellow light and the expectation that it will turn red soon, precaution is required. The red light takes priority, and the vehicle must prepare to stop. Keeping the speed low as a precaution at this specific moment is appropriate, ... The vehicle on the front left is moving away, so it does not pose an immediate threat.

## Decision

DC

Figure 17: Failure case that running the red light.

## Scene Description

There are 2 important objects in the scene. The information of these key objects are: <ref> In 'CAM_FRONT', vehicle is located approximately 16.975 m to the front left of the ego car is driving in the same direction </ref> <box> (283, 432), (396, 584) </box> <ref> In 'CAM_FRONT', vehicle is located approximately 15.945 m to the front of the ego car is drive in the same direction </ref> <box> (441, 476), (528, 587) </box>.

## Ego State

Your current speed is 1.2176832130196866 m/s. Your current steering value is 0.0.

## Reasoning

Based on the provided information, we understand that there is a vehicle about 15.945 meters directly in front of us, driving in the same direction, and this vehicle is within the bounds of the ego lane, which suggests it is in our direct path.
Our current speed is low (1.2176832130196866 m/s), but since there's an object in close proximity (under 25m) on the ego lane, our priority is to adjust our speed accordingly to ensure safety. Our steering value is 0, indicating that we are moving straight ahead, and thus the vehicle directly in front of us is relevant to our lane. The distance to the vehicle ahead is greater than 15 meters, so we don't need to come to a complete stop yet, but we should decrease speed.

## Decision

DC

Figure 18: Failure case of collisions caused by other vehicles.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The motivations and contributions are well depicted and summarized in the abstract and introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: No theoretical results are included.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The data preparation and implementation details are clearly and fully presented in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We provide open access to the project page, which includes the GitHub repository for subsequent releases.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All details are presented in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have provided the mean value and standard deviation of several runs for the main results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The information on the computer resources is illustrated in the implementation details of Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We make sure that the research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed the potential societal impacts in Section 6.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited the original paper that produced the dataset.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.