DOLPHINS: MULTIMODAL LANGUAGE MODEL FOR DRIVING

Yingzi Ma[1] Yulong Cao[2] Jiachen Sun[3] Marco Pavone[2,4] Chaowei Xiao[1,2]

[1]University of Wisconsin-Madison [2]NVIDIA
[3]University of Michigan [4]Stanford University

ABSTRACT

The quest for fully autonomous vehicles (AVs) capable of navigating complex real-world scenarios with human-like understanding and responsiveness. In this paper, we introduce `Dolphins`, a novel vision-language model architected to imbibe human-like abilities as a conversational driving assistant. `Dolphins` is adept at processing multimodal inputs comprising video (or image) data, text instructions, and historical control signals to generate informed outputs corresponding to the provided instructions. Building upon the open-sourced pretrained Vision-Language Model, OpenFlamingo, we first enhance `Dolphins`'s reasoning capabilities through an innovative Grounded Chain of Thought (GCoT) process. Then we tailored `Dolphins` to the driving domain by constructing driving-specific instruction data and conducting instruction tuning. Through the utilization of the BDD-X dataset, we designed and consolidated four distinct AV tasks into `Dolphins` to foster a holistic understanding of intricate driving scenarios. As a result, the distinctive features of `Dolphins` are characterised into two dimensions: (1) the ability to provide a comprehensive understanding of complex and long-tailed open-world driving scenarios and solve a spectrum of AV tasks, and (2) the emergence of human-like capabilities including gradient-free instant adaptation via in-context learning and error recovery via reflection. See the project page for demo, examples, and request pre-trained models: https://vlm-driver.github.io/.

Keywords Autonomous Driving · Vision Language Model

1 Introduction

The odyssey toward achieving full autonomy in vehicular systems has been a crucible of innovation, melding insights from artificial intelligence [1], robotics [2], and automotive engineering [3]. The essential aspiration is to design autonomous vehicles (AVs) capable of maneuvering through complex real-world driving situations with human-like understanding and responsiveness.

Current autonomous driving systems (ADS) [4] are data-driven and typically modular, dividing tasks like perception, prediction, planning and control [5]. However, these systems struggle with integration and performance in varied situations. End-to-end (E2E) designs offer a direct sensory input to control output mapping, but they lack interpretability, posing challenges in safety and regulatory compliance [6, 7, 8].

Moreover, existing ADS exhibit many limitations when compared with human drivers including: (1) **Holistic Understanding and Interpretation**: existing data-driven Autonomous Driving Systems (ADS) often fall short in holistically understanding and interpreting dynamic and complex scenarios, especially those within the long-tail distribution of open-world driving environments [9, 10]. For instance, considering a scenario where a ball bounces onto the road, followed by a child running after it, a human driver could immediately deduce the potential danger and act accordingly to prevent any mishap, leveraging a blend of common sense, past experiences, and a fundamental understanding of human behaviors. In contrast, existing ADS might struggle to interpret this scenario accurately without prior exposure to a large amount of similar data. This lack of holistic understanding limits the system's ability to generalize well across unexpected scenarios that may be located in the long tail of the data distribution [11, 12]. (2) **Instant Learning and Adaptation**: unlike human drivers who can instantly learn and adapt to new scenarios with just a few examples, existing ADS requires extensive training with large amounts of data to handle new situations. For example, a human

driver can quickly learn to navigate around a new type of road obstacle after encountering it once or twice, whereas an ADS might require exposure to many similar scenarios to learn the same lesson. (3) **Reflection and Error Recovery**: existing ADS typically employ feedforward processing during operation, lacking the capability for real-time correction based on feedback and guidance. In contrast, human drivers can correct their driving behavior in real time based on feedback. For instance, if a human driver takes a wrong turn, they can quickly adjust their decision based on the error feedback, whereas an ADS might struggle to quickly recover from the error feedback [13, 14].

These limitations underline the need for an intermediate framework that can bridge the gap between the current state of AV systems and human-like driving. Recent advancements in (multimodal) large language models (LLMs) [15, 16, 17] with emergent abilities offer a hopeful path toward addressing these challenges. These models are endowed with a rich repository of human knowledge, laying the foundation for valuable insights that could significantly improve ADS. However, these model are mainly trained on general vision and language data, which restricts their efficacy in the specialized driving domain. Moreover, current model designs can only digest static image and text data to generate zero-shot decisions, lacking in handling temporal video input and in-context learning.

In this paper, we propose `Dolphins` (shown in Figure 1), a vision language model (VLM) specifically tailored for AVs, as a **conversational driving assistant** to help reduce the gap between existing ADS and human-like driving.

Built upon OpenFlamingo [18], `Dolphins` is adapted to the driving domain through a series of specialized instruction datasets and targeted instruction tuning. We first build an image instruction-following dataset with grounded CoT responses based on some public VQA datasets [19, 20, 21, 22], visual instruction datasets [15, 23], and ChatGPT, to ground the fine-grained reasoning capability into OpenFlamingo models. Then, we utilize BDD-X [24] to establish our instruction dataset, focusing on four key AV tasks: behavior comprehension, control signal forecasting, behavior analysis, and in-depth conversation.

`Dolphins` demonstrates an advanced understanding of complex driving scenarios and human-like abilities such as instant learning, adaptation, reflection, and reasoning, which significantly reduces the gap between existing ADS and human-like driving. Notably, `Dolphins` showcases broad task applicability across perception, prediction, and planning, thanks to its comprehensive scenario understanding. It interprets static and dynamic scenarios, integrates environmental factors, and handles downstream prediction and planning tasks effectively.

Furthermore, `Dolphins`'s in-context learning ability allows it to quickly adapt to new driving conditions, a significant advancement over existing models. Its error recovery mechanism enhances model accuracy and reliability, making it a valuable tool for real-world driving scenarios. Importantly, `Dolphins` offers interpretability, a crucial factor in building trust and ensuring transparency in ADS operations.

We summarize our contribution as three folds:

- We propose a VLM-based conversational driving assistant, `Dolphins`, that plans high-level behaviors like humans complementary to ADS.
- We have devised a Grounded Chain of Thought (GCoT) process to initially endow `Dolphins` with the capability of Chain of Thought reasoning. Following this, we align the model with AV tasks, facilitating its understanding of the AV context despite the limited scope of the available dataset. This approach not only compensates for dataset constraints but also enables `Dolphins` to effectively decompose complex tasks and learn the underlying subtasks.
- We demonstrate the prominent capability of `Dolphins` spanning scene understanding and reasoning, instant learning and adaptation, and reflection and error recovery, with both quantitative metrics and qualitative demonstrations.

2 Related Work

Autonomous Driving with LLMs The recent wave of research focuses on utilizing Large Language Models (LLMs) as the driving agents to address autonomous driving-related tasks, such as perception, reasoning, planning, and other related tasks. For instance, DriveLikeHuman [25] designs a new paradigm to mimic the process of human learning to drive based on LLMs while GPT-Driver [26] leverages GPT-3.5 to assist autonomous driving in dependable motion planning. In a parallel vein, SurrealDriver [27] uses the CARLA simulator for building a LLM-based DriverAgent with memory modules, including short-term memory, long-term guidelines, and safety criteria, which can simulate human driving behavior to understand driving scenarios, decision-making, and executing safe actions. DriveLM [28] and NuPrompt [29] introduce innovative driving tasks based on the NuScenes dataset [30]. Specifically, DriveLM leverages the idea of graph-of-thought (GoT) to connect graph-style QA pairs for making decisions and ensuring explainable planning using the powerful reasoning capabilities of LLMs for autonomous driving. NuPrompt employs

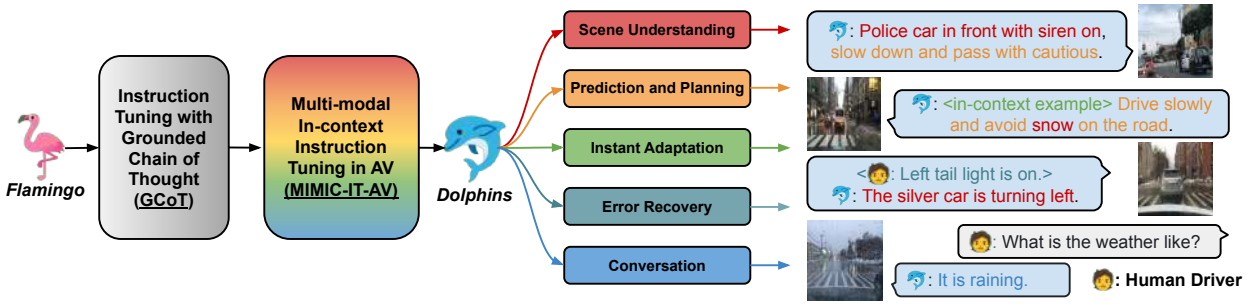

Figure 1: `Dolphins` **overview.** Demonstrations in Section 5 show that `Dolphins`'s capabilities on a group of subtasks belonging to the two dimensions of **holistic understanding and reasoning**, and **human-like capabilities**. The former encompasses autonomous driving-related capabilities such as scene understanding and prediction and planning for the ego car's behavior. The latter analyzes three human-level abilities: rapid learning and adaptation, reflection and error recovery, and interactive conversation.

LLMs to formulate a new prompt-based driving task that focuses on object tracking. However, these works only accept linguistic input and lack the incorporation of rich visual features. In contrast, `Dolphins` excels as a cohesive large vision-language model, not only possessing the reasoning and planning capabilities of LLMs but also exhibiting proficiency in understanding diverse visual features.

Large Vision-Language Models (LVLMs). Progress has been witnessed in employing the powerful capabilities of large language models like LLaMAs [31, 32], Vicuna [33], and MPT [34] to enhance Large Vision Language Models (LVLMs), such as Flamingo [35] and BLIP-2 [36]. Recently, to unlock the capabilities of LVLMs to align with human preferences, LLaVA [15] and MiniGPT-4 [37] pioneers visual instruction tuning, with subsequent efforts like Otter [38], InstructBLIP [39], and Mplug-owl [40] following suit. Building upon these significant contributions, the potential of LVLMs has been progressively realized, leading to their swift adaptation across diverse domains, including video chatting [41, 42, 43, 44], embodied AI [45, 46, 47, 48], 3D-world understanding [49, 50], medical healthcare [51, 17, 52], marine sector [53], etc. Inspired by these models, we propose `Dolphins`, the vision-language model designed to facilitate autonomous driving-related interaction. This field has already witnessed some related research efforts, such as HiLM-D [54], DriveGPT4 [55], LINGO-1[1]. However, DriveGPT4 and HiLM-D exhibit a deficiency in task diversity and only accept one video as input, which can significantly curtail the LVLMs' ability to generalize to unseen instructions. To mitigate this issue, we propose `Dolphins`, which is extended from OpenFlamingo [18] with strong in-context learning capabilities. Furthermore, we employ in-context instruction tuning [56] to enhance few-shot adaptations of our model. Consequently, `Dolphins` is proficient in handling diverse video inputs and exhibits the capacity for rapid adaptation to unseen instructions through in-context learning.

Multimodal In-context Learning. Flamingo [35] is the pioneering work to support in-context learning in the multi-modal domain by constructing MultiModal MassiveWeb(M2W) and employing the upstream training. Following this line of thought, the other works [38, 57] focus on constructing text-image interleaved instruction datasets by adding related in-context exemplars, thus enhancing the instruction comprehension ability of MLLMs while preserving the in-context learning capacity.

3 Method

To equip VLMs with a comprehensive understanding and human-like capabilities, we need to ground them within the autonomous vehicle (AV) context to support a variety of tasks. However, limited task-specific labeled data in AV has posed a challenge for such grounding. To address this, we initially foster comprehensive reasoning in VLMs by utilizing chain-of-thought (CoT) principles [46], applied to a custom VQA dataset. Specifically, we designed a video-text interleaved dataset by enriching existing datasets, covering all functionalities at a coarse level. Tuning the VLM on this dataset enables it to develop capabilities for handling tasks with finer granularity.

We introduce our methodology in this section. First, we describe our grounding method for CoT in § 3.1. Next, we elaborate on the creation of our video-text interleaved dataset for autonomous driving in § 3.2 along with our devised tasks. Finally, we detail the multi-modal in-context instruction tuning for AD in § 3.3.

[1] https://wayve.ai/thinking/lingo-natural-language-autonomous-driving/

3.1 GCoT Instruction tuning

Reasoning abilities based on fine-grained understanding are essential in AD. This is because the model needs to perceive the spatial information of objects in the perceived visual input to infer their relationships and interactions with the ego vehicles. To the best of our knowledge, most VLMs in the literature lack fine-grained multimodal understanding of the visual modality (e.g., image and video), primarily due to their coarse-grained alignment in vision-language pre-training [58, 59]. Although HiLM-D [54] delivers a fine-grained understanding capabilities of VLMs by feeding high-resolution images and adding a detection module in autonomous driving (AD), it is restricted by the quality of the existing datasets. To further improve the fine-grained understanding of VLMs, we devise grounded CoT (GCoT) instruction tuning and develop a dataset that grounds this ability.

Ideally, GCoT capability should naturally occur within the autonomous vehicle (AV) context by utilizing datasets comprised of massive driving videos paired with relevant question and answer sets. However, the availability of such datasets in the AV domain is markedly limited and it is hard to capture the spatial information of a driving video. We, therefore, design a new method to circumvent this limitation. Specifically, we initially ground the GCoT capability in a general image dataset. Recognizing the proficiency of ChatGPT in demonstrating reasoning ability through detailed step-by-step reasoning, we define a general pipeline for generating GCoT response using ChatGPT to enrich the current VQA datasets. As shown in Figure 2, this process is divided into three steps: **(1)** briefly describe the content of the image. **(2)** identify the object in the question and describe its spatial position. **(3)** if the question requires reasoning, provide the reasoning process in this step. Finally, we combine the sentences generated by ChatGPT in these three steps and append "So the answer is {answer}" at the end to form a complete GCoT response. This approach involves training the model on diverse visual data with GCoT response, where it learns to articulate its reasoning process in a step-by-step manner for various scenarios and objects that might not be specific to driving but are crucial for building foundational reasoning skills. Detailed information can be found in Appendix A.

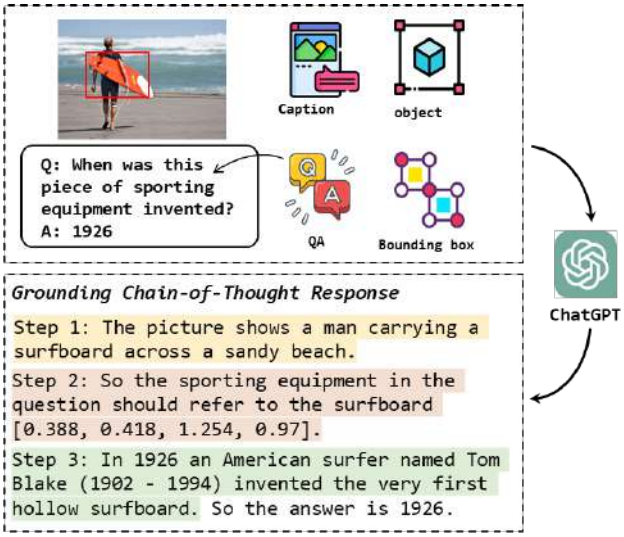

Figure 2: The process of generating GCoT response for VQA tasks to enhance the fined-grained reasoning capability of VLMs. ChatGPT is prompted to generate GCoT step by step from text input.

Subsequently, we transition this grounded capability to the AV context. This transfer involves aligning the model with AV-specific scenarios, where it applies the generalized reasoning ability to the nuanced and dynamic environment of autonomous driving. The transfer process includes fine-tuning the model on AV-specific datasets, which, although limited, contain critical driving scenarios, road conditions, and interactions. This stage focuses on adapting the general reasoning skills to the specialized requirements of AV scenarios, ensuring that the VLM can apply its fined-grained reasoning capability to real-world driving situations effectively.

In summary, the development of the fine-grained capability in our VLM is a multi-stage process. It begins with grounding the model in a general image dataset with GCoT responses generated by ChatGPT, followed by a careful transfer and fine-tuning of this skill in the specific context of AD. The use of both real and synthetic AV datasets ensures a comprehensive and robust training regime, preparing the VLM to handle the intricate and varied challenges of autonomous vehicular navigation with nuanced, step-by-step reasoning. nstructions with just a handful of annotated examples in autonomous driving-related tasks.

3.2 Devised Instruction Tasks for Autonomous Driving

For autonomous driving-related video understanding, we include four tasks critical for perception, prediction and planning as shown in Figure 3: **(1) Behavior Understanding.** For predicting action description labels in the BDD-X dataset, we employ the same instructions for description (noted as $\mathcal{Q}_a$) from DriveGPT4 [55] to guide the model in learning the ego car behavior in videos. **(2) Behavior Reasoning.** Similar to the Behavior Understanding task, we also utilize instructions of justification (noted as $\mathcal{Q}_j$) from DriveGPT4 to enable the model to interpret the behavior of the ego car. **(3) Prediction with Control Signals.** In the BDD-X dataset, the time durations of different video segments vary. Hence, in this task, the number of historical control signals provided depends on the duration of the

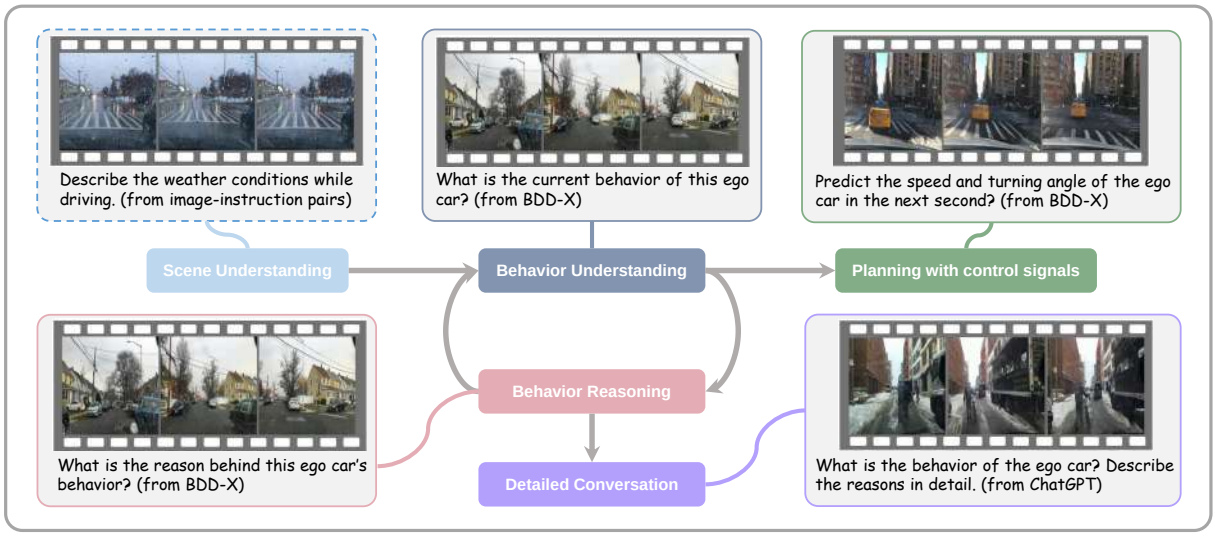

Figure 3: Overview of our proposed dataset. Compared with the previous datasets, we employ RICES (Retrieval-based In-Context Example Selection) [60] approach to choose in-context examples for each sample. Additionally, We introduce the "Detailed Conversation" task to train our model to generate detailed responses that align closely with human-preferred responses. This instruction is aimed at unlocking the latent potential of the foundation model, which has instruction fine-tuned on the dataset consisting of image-instruction-response triplets.

video segments. VLMs are required to predict the ego car's speed and turn angle for the next second based on these control signals (e.g., speed, accelerator, and turn angle). **(4) Detailed Conversation.** The three tasks above tend to lean towards traditional vision-language tasks (short answer). Consequently, we aim to introduce more detailed conversations to enhance instruction generalization ability for human-preferred responses (long answer). Specifically, we rely on the in-context learning ability of ChatGPT [61] to enrich the action description and reasoning labels for generating human-preferred responses in terms of traffic rules, potential risks of the behavior, driving precautions, etc.

To construct a dataset suitable for end-to-end autonomous driving systems, we collect video segments and labels sourced from the BDD-X dataset [24]. The BDD-X dataset comprises roughly 7,000 videos, with each video being subdivided into multiple segments, each of which conveys distinct behaviors of the ego car along with corresponding textual annotations. There are approximately 25,000 examples in total, with annotations including action descriptions (e.g., "the car stops") and action reasoning (e.g., "because the traffic light is red"). Following the previous work [55], we leverage the BDD-X dataset to develop our visual instruction-following dataset for autonomous driving, consisting of four distinct autonomous driving-related tasks and their corresponding instructions. However, due to limitations in the diversity of tasks and instructions, the VLM trained on this dataset exhibits a significant deficiency in its ability of zero-shot generalization to unseen tasks. Thus, we leverage multi-modal in-context instruction tuning [38] to assist our model in the rapid adaptation to new instructions with just a handful of annotated examples in autonomous driving-related tasks.

Integrated with our devised tasks, our proposed dataset comprises 32k video-instruction-answer triplets, with 11k of them belonging to the detailed conversation task generated by ChatGPT. The remaining three tasks collectively contain 21k triplets from labels of the BDD-X dataset. Noticed that the proposed tasks for constructing the dataset are a coarse-grained set that can be resolved better by a CoT process. As a result, the model grounded on CoT will be forced to emerge diverse capabilities beyond such tasks in order to achieve good results on the constructed dataset during the instruction tuning process.

3.3 Multi-modal In-context Instruction Tuning in Autonomous Driving

In the NLP community, training models with in-context examples is widely considered beneficial for facilitating the model's capacity to learn new tasks from several input-output examples, known as few-shot prompting [57, 62, 63, 64, 65].In terms of visual instruction tuning, Otter [38] has introduced in-context instruction tuning to preserve the VLM's few-shot, in-context learning capabilities. Motivated by these notable works, we introduce in-context instruction tuning to the autonomous driving domain. This field currently faces a severe shortage of diverse instruction-following datasets.

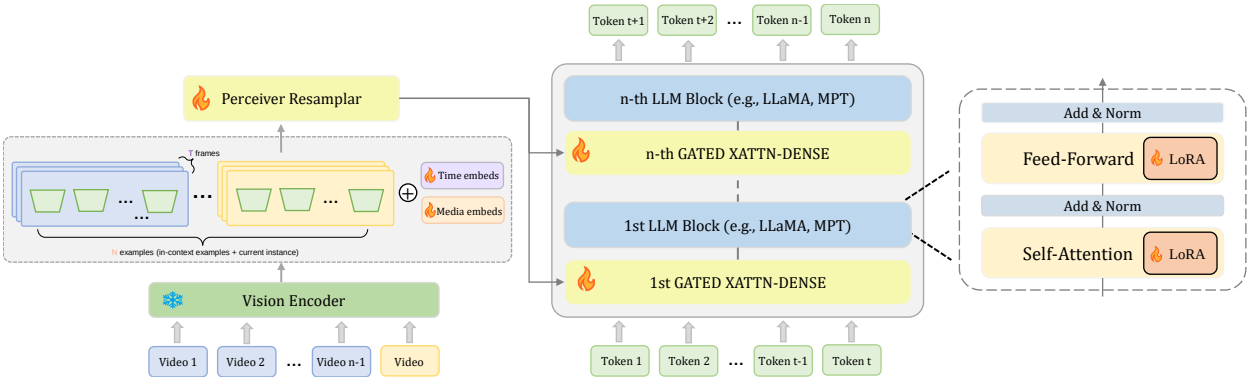

Figure 4: `Dolphins`'s model architecture.

We aim to enhance VLMs' in-context learning capabilities to facilitate the generalization of models across a spectrum of autonomous driving-related tasks.

In pursuit of the aforementioned objective, we employ OpenFlamingo [18] as our foundational VLM. OpenFlamingo, a reimplementation of Flamingo [35], is trained on the integration of image-text interleaved Lion-2B [66] and MMC4 [67] datasets to enhance its in-context learning capabilities. Our autonomous driving-related instruction dataset, as described in Section 3.2, adopts a format comprising video-instruction-answer triplets. Consequently, we employ a retrieval approach to select in-context examples for each triplet. Specifically, we utilize `VideoMAE` [68] and `text-embedding-ada-002` [2] as the image encoder $\mathbf{E}_{\texttt{Image}}$ and text encoder $\mathbf{E}_{\texttt{Text}}$, which map a video segment $\mathbf{X}_v$ or a text (instruction-answer pairs) instance $\mathbf{X}_t$ to a d-dimensional latent space. Then, we subsequently retrieve in-context examples based on the cosine similarity of their representations for each sample $\mathbf{Z}^i = (\mathbf{X}_v^i, \mathbf{X}_t^i)$. We denote this retrieval pipeline as $\mathcal{R}$:

$$\mathcal{R}(\mathbf{Z}^i) = \{\underset{\mathbf{X}_v}{\text{Top}\, k} \left(\cos(\mathbf{E}_{\texttt{Image}}(\mathbf{X}_v^i), \mathbf{E}_{\texttt{Image}}(\mathbf{X}_v))\right), \tag{1}$$

$$\underset{\mathbf{X}_t}{\text{Top}\, k} \left(\cos(\mathbf{E}_{\texttt{Text}}(\mathbf{X}_t^i), \mathbf{E}_{\texttt{Text}}(\mathbf{X}_t))\right)\} \tag{2}$$

$$= \{\hat{\mathbf{Z}}^1, \ldots, \hat{\mathbf{Z}}^{2k}\}. \tag{3}$$

Where k represents that we respectively search k nearest samples in both text-encoded and image-encoded latent space. In essence, examples featuring behaviors akin to those of the ego car within the video are more likely to be selected. In our previous research endeavors [69], we observed that the retrieval of in-context examples based on textual similarity proved more effective in preserving the VLM's in-context learning ability compared to using image features. We posit that this conclusion is equally applicable to video-text pairs. Therefore, we only utilize in-context examples retrieved by text embedding similarity and constrain the provision of in-context examples to a maximum of $k = 3$ per triplet during the training stage.

4 Training

4.1 Model Architecture

Our model is based on OpenFlamingo architecture, named `Dolphins`. The model consists of a vision encoder from CLIP [70], a perceiver resampler to receive the visual features from the vision encoder, and a text encoder from large language models (e.g., LLaMA [16], MPT [34]) equipped with gated cross-attention layers for image-text interactions. However, unlike Flamingo, OpenFlamingo lacks the capability to support video inputs. Therefore, to mitigate the vanishing of global temporal features resulting from the aggregation of spatial features, we introduce a set of learned latent vectors as temporal position embeddings. Similarly, another set of learned latent vectors is incorporated to function as media position embeddings, introducing essential ordering information within the few-shot prompt. The inclusion of these embeddings has led to a noteworthy enhancement in the model's ability in video understanding. To preserve the pretraining knowledge and reduce computing consumption, We freeze both the encoders and only finetune the perceiver resampler module, gated cross-attention layers, and LoRA [71] module added to the text encoder, as shown in Figure 4.

[2] https://platform.openai.com/docs/guides/embeddings/what-are-embeddings

Definition: [Task Definition] (four autonomous driving-related tasks)
 # in-context exemplars
User: <image>is a driving video. [instruction]
GPT: <answer> [answer] <endofchunk>
...
User: <image> is a driving video. [instruction]
GPT: <answer> [answer] <endofchunk>
 # current instance
User: <image> is a driving video. [instruction]
GPT: <answer> [answer] <endofchunk>

Table 1: <image> and <endofchunk> tokens are originally from the OpenFlamingo training paradigm, and we follow Otter to include a new token <answer> for intercepting the target answer of the model output more easily. Note that only green sequence/tokens are used to compute the loss and we train our model using a cross-entropy loss.

4.2 Implementation Details

Inspired by Otter, we employ a similar format to prepare our instruction-tuning data. Additionally, we also introduce a specific task definition at the beginning of each task as a task-level instruction, which aids the model in comprehending the broader context of autonomous driving-related video-instruction pairs of the same type. The training data is structured as shown in Table 1.

In contrast to the existing video-related VLMs [72, 73, 74, 75, 55], which typically employ a two-stage training framework involving a first stage for aligning video-text features followed by a second stage for visual instruction tuning, we remove video alignment stage on general video-text pairs datasets and instead fine-tune initially on image instruction-following datasets that we collect, equipped with grounded CoT templates to enhance fine-grained understanding and reasoning abilities. Subsequently, we further fine-tuned using our proposed autonomous driving instruction dataset to transfer the model's capabilities from images to autonomous vehicles (AVs).

To optimize Dolphins, We utilize DeepSpeed [76] for optimization during the training process. An AdamW [77] optimizer is used, with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and a weight decay of 0.01. All training runs on 4 NVIDIA A100 GPUs, with a total batch size of 128, a learning rate of 2×10^{-5} for the second stage. The maximum sequence length is fixed at 1024 and BF16 precision is used for both training and inference.

5 Demonstration

In this section, we show various examples of demonstrations of Dolphins from two dimensions (shown in Figure **??**): holistic understanding and human-like capabilities, on both zero- and few-shot settings. We will first summarize key desiderata of AV tasks tailored for the VLM setup (§ 5.1). Then we will show that Dolphins has a holistic understanding with emerged capabilities accomplishing these diverse tasks spanning perception (§ 5.1.1), prediction, and planning (§ 5.1.2) even for unseen instructions. Also, we will show the human-like capabilities of Dolphins in (1) rapid learning and adaptation through in-context learning (§ 5.2.1); (2) error recovering through reflection (§ 5.2.3); and (3) communicating with human through interactive conversation (§ 5.2.2).

5.1 Holistic Understanding and Reasoning

We transform the traditional perception, prediction, and planning design into a group of subtasks tailored to the advantages of VLM in terms of open-vocabulary detection and comprehensive semantic reasoning.

- **Attributed road agents/traffic elements**. Compared to bounding boxes and tracked history for road agents categorized in a close set of labels, Dolphins should be able to understand road agents and traffic elements with comprehensive semantic attributes including: an open vocabulary semantic type (e.g., a police vehicle, a kid pedestrian, etc.); a semantic status (e.g., with right turn light on, with green light on, etc.); a behavior description if it is a dynamic road agent (e.g., turning right in slow speed, parallel parking, etc.). These comprehensive attributes are crucial for understanding the rationale behind the scene with VLM (e.g., giving road to a police vehicle with siren on, right turn light on infering a right turn behavior, etc.).
- **Operational design domain (ODD)**. Dolphins should be able to extract ODDs including weather, time of day, and geolocation. ODDs provide a high-level driving concept that supervises the downstream prediction and planning strategy (e.g., driving slowly and keeping longer safety distance on snowy days).

- **Ego agent behavior**. Dolphins should be able to understand the behavior of the ego agent (e.g., the ego vehicle is turning right). With an ego-centric video as the input, it is crucial for the model to understand the ego agent behavior first to condition other road agents behavior.
- **Predicted road agent behavior**. Compared to prediction-based on trajectory, Dolphins should be able to provide a (probalistical) behavior prediction of road agents' behaviors to cover different modes in the future (e.g., the vehicle on the front right side could follow the lane or change lane to the front of the ego). This is crucial for the ego vehicle to understand the intentions of other road agents and plan its corresponding reactive behavior in advance.
- **Ego agent future plan**. Dolphins should be able to reason on top of the perceived scene and provide a instruction (e.g., as the vehicle on the front right could change lanes to the front of the ego, the ego should drive with caution and be ready to yield to it). By featuring a reasoning of things not to do and contingency planning, Dolphins is capable of planning for safe and flexible actions foreseeing different modes of other road agents.

In the following section, we will demonstrate Dolphins's capability on such subtasks spanning over perception, prediction, and planning through holistic understanding and reasoning.

5.1.1 Perception

In evaluating Dolphins's holistic understanding on perception tasks, we focus on the **understanding** of scenario and behavior. These competencies are pivotal for autonomous systems, necessitating acute recognition and comprehension of environmental and situational nuances. Our demonstrations reveal Dolphins's capacity to interpret driving-related visual content, spanning subtasks described in § 5.1:

- **Semantic attributes of road agents & traffic elements.** Dolphins is able to capture various types of road agents and traffic elements with attributes (e.g., black car, red traffic light, evident in Figures 5, 7, 8 and 9);
- **ODDs.** Dolphins is able to understand different ODDs such as weather conditions (as depicted in Figures 7 and 8), and illumination (as shown in Figures 5 and 10);
- **Traffic conditions.** Dolphins is proficient in pinpointing the precise driving location (as observed in Figure 5), and overall traffic status (as observed in Figure 5, 7, and 8).
- **Behavior of road agents.** Dolphins is able to understand behaviors of road agents (as shown in Figure 6, 9, and 10).
- **Ego agent behavior.** Dolphins is able to understand the ego agent behavior by inferring from the ego-centric video (as detailed in Figures 8, 6, 9, and 10).

This comprehensive perceptual insight allows for a high-quality and fluent natural language response from the system, encompassing a wide spectrum of capabilities crucial for autonomous navigation.

5.1.2 Prediction and Planning

Following DriveLM [28], We also evaluate Dolphins's prediction and planning capabilities, which involve utilizing the **reasoning** ability of VLMs to assist the driver in making decisions and ensuring explainable planning. In Figures 12, 13, 19 and 15, we showcase our model's multimodal ability to predict the behavior of other vehicles in the future and determine whether these vehicles affect the ego agent's trajectory. Figures 14, 18, 16, and 17 present some examples that demonstrate Dolphins can generate comprehensive plans for the ego car based on current traffic conditions. Furthermore, Figures 19 and 15 demonstrate Dolphins's capabilities to make reasonable and safe plans based on the contingent behavior of other agents, which is considered crucial in real-world driving scenarios. However, due to the lack of relevant instructions during training, we currently recommend using in-context learning ability, ongoing dialogue, and control signals to assist Dolphins in completing these two tasks. We believe this is still an under-explored facet and we are working on it.

5.2 Human-like Capabilities

In this set of demonstrations, we will show the human-like capabilities of Dolphins in (1) rapid learning and adaptation through in-context learning (§ 5.2.1); (2) error recovering through reflection (§ 5.2.3); and (3) interpretability through interactive conversation (§ 5.2.2).

5.2.1 Rapid Learning and Adaptation

In this demonstration, we document the agility of Dolphins to rapidly assimilate and adapt to new driving conditions—a process akin to human learning. This facet is examined by presenting Dolphins with a series of unforeseen scenarios and monitoring its response efficiency and accuracy after exposure to a limited set of examples. The tasks are designed to test the model's in-context learning capabilities by progressively introducing more complex and previously unseen driving scenarios. Through this, Dolphins demonstrates its ability to leverage prior knowledge and quickly adapt, suggesting that even in the absence of extensive pre-training on certain tasks, it can still formulate accurate predictions and actions using a sparse sampling of in-context examples. Specifically, Figure 20 shows our model's ability to learn certain common-sense knowledge from in-context examples, such as "*You cannot determine the current time by the light in a tunnel.*". Besides, Dolphins demonstrates the excellent ability to respond to unseen instructions in various styles, such as "*What if*" and "*What are you doing*", as shown in Figures 20 and 21.

5.2.2 Interactive Conversation

In this demonstration, we subject Dolphins to an evaluation of its conversational skills through multi-turn dialogues, gauging its competence in engaging with human drivers under varying conditions. Utilizing a set of instructions primarily derived from LINGO-1 [3], we present Dolphins with a spectrum of queries reflective of real-world driving interactions. As shown in Figures 22, 23, and 24, the conversations are constructed to assess Dolphins's ability to comprehend and respond to nuanced language, maintain context over multiple exchanges, and offer informative and contextually relevant responses spanning from potential hazards in the scene to ego planning and the reasoning behind the scene. The results from these interactions indicate that Dolphins possesses a robust conversational ability, distinguishing itself significantly from other contemporary driving-related Vision Language Models in terms of linguistic flexibility and contextual understanding. In the future, this could be a foundation for a human interface that builds up trust between AV and road users or its passengers.

5.2.3 Reflection and Error Recovering

This demonstration is devoted to showcasing Dolphins's self-assessment and error correction mechanisms. We present instances where the model first produces a suboptimal response to a given driving scenario and is subsequently provided with feedback. The focus here is on how Dolphins reflects on this feedback to identify and correct its mistakes. We evaluate the effectiveness of these mechanisms through a series of iterative interactions, illustrating the model's capacity for reflection, error identification, and the implementation of corrective measures. The results (evidenced in Figure 25, 26 and 27) underscore the model's ability to not just detect and recover from errors in a manner that mirrors human cognitive processes, but to also refine its subsequent responses, thereby enhancing overall performance and reliability. For example, Figure 25 presents that Dolphins is able to reflect on the tail-light signals and predict the correct right turning behavior of the black car; and Figure 26 shows that Dolphins can reflect and revise the reasoning for the unnecessary yielding decision when encountering a car in a narrow street.

6 Conclusion and Future Directions

As we conclude our exploration into Dolphins, a novel vision-language model designed for enhancing autonomous vehicles (AVs), we reflect on the significant strides made and the challenges ahead. Dolphins has demonstrated a remarkable capacity for holistic understanding and human-like reasoning in complex driving scenarios, marking a substantial advancement in the realm of autonomous driving technology. By leveraging multimodal inputs and employing the innovative Grounded Chain of Thought (GCoT) process, Dolphins has shown its proficiency as a conversational driving assistant, capable of addressing a wide spectrum of AV tasks with enhanced interpretability and rapid adaptation capabilities.

However, our journey towards fully optimizing Dolphins for real-world application in AVs encounters notable challenges, particularly in computational overhead and feasibility. Our assessment of Dolphins's performance on the DriveLM dataset, a realistic benchmark for real-world driving scenarios, revealed an average inference time of 1.34 seconds on an NVIDIA A100, indicating a potential limitation in achieving high frame rates on edge devices. Additionally, the power consumption associated with running such sophisticated models in vehicles presents a significant hurdle for deployment. These findings underscore the necessity of further advancements in model efficiency. Looking forward, the development of customized and distilled versions of these models, as suggested by emerging research [78], appears to be a promising direction. These streamlined models are anticipated to be more feasible for deployment on edge devices, balancing computational demands with power efficiency. We believe that continued exploration and

[3] https://wayve.ai/thinking/lingo-natural-language-autonomous-driving/

innovation in this domain are vital for realizing the full potential of AVs equipped with advanced AI capabilities like those offered by Dolphins.

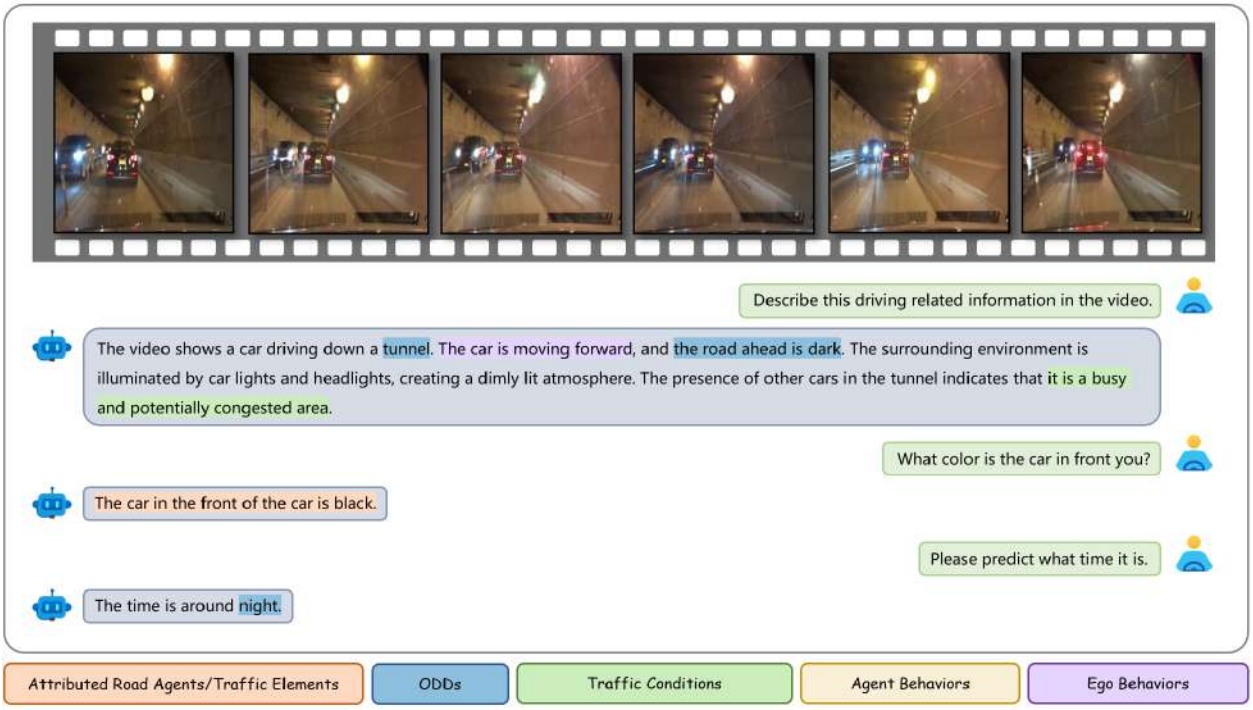

Figure 5: An example showcasing Dolphins's capability in **scenario understanding** § 5.1.1. The video features an ego car driving in a tunnel. Dolphins can identify the environment in which the ego car is situated and accurately determine the color of the front vehicle as well as infer the current time.

Figure 6: An example showcasing Dolphins's capability in **scenario understanding** and **behavior understanding** § 5.1.1. The video features an ego car making a right turn. Dolphins can identify these contents. ~~Words~~ means hallucination.

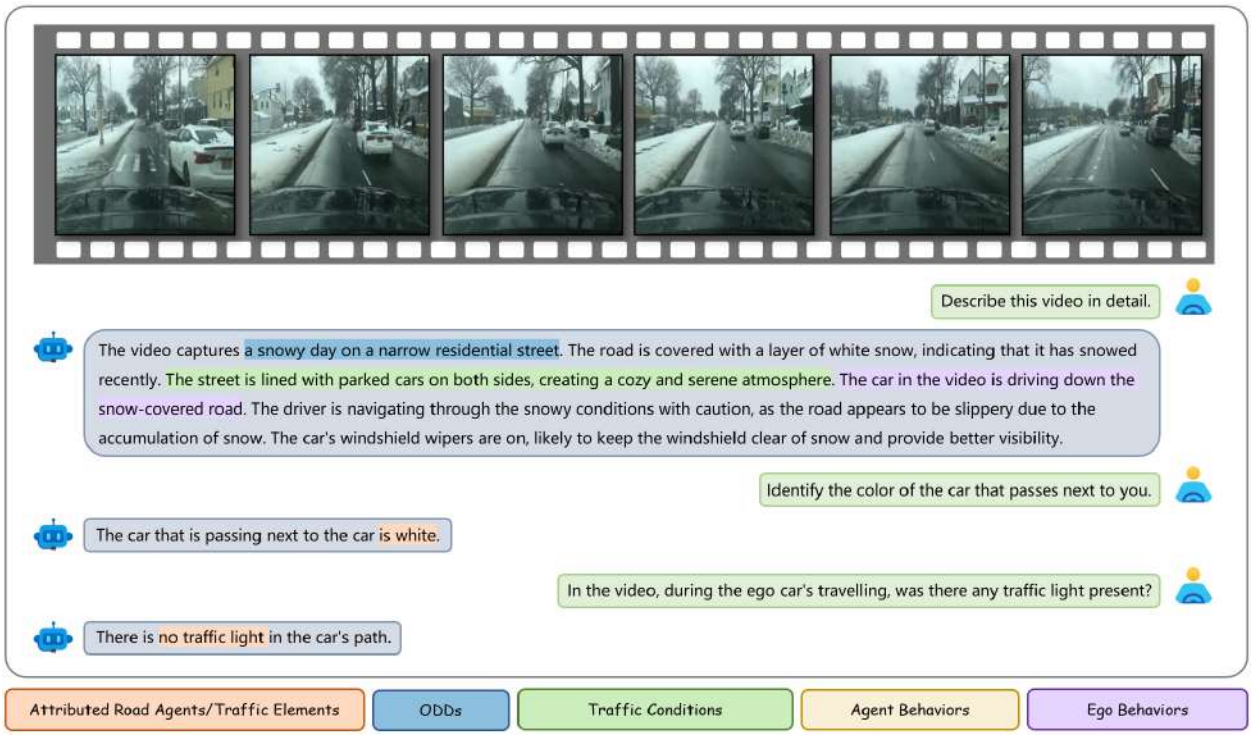

Figure 7: An example showcasing `Dolphins`'s capability in **scenario understanding** § 5.1.1. The video features an ego car driving on a snowy street. `Dolphins` can identify the environment in which the ego car is situated, the presence of the traffic light, and accurately determine the color of the passing vehicle.

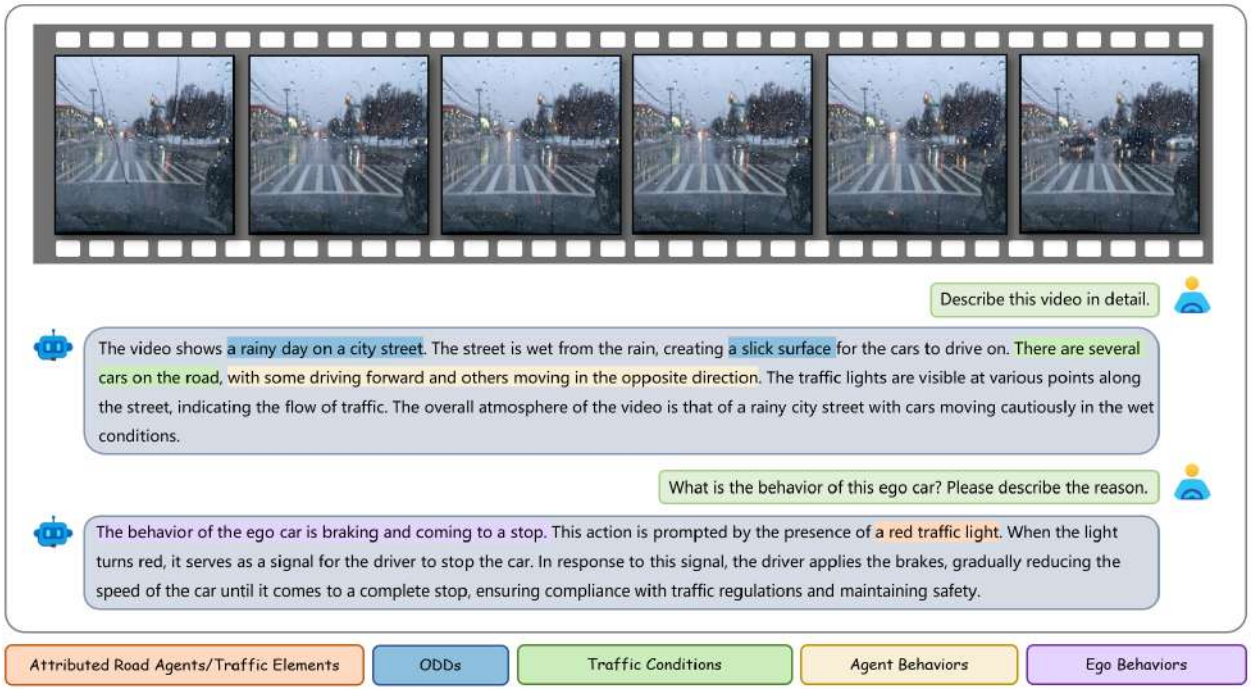

Figure 8: An example showcasing `Dolphins`'s capability in **scenario understanding** and **behavior understanding** § 5.1.1. The video features an ego car stopping at an intersection on a rainy day. `Dolphins` comprehensively describes the environment in which the ego car is situated, the behavior of the ego car, and can infer the reasons for its behavior.

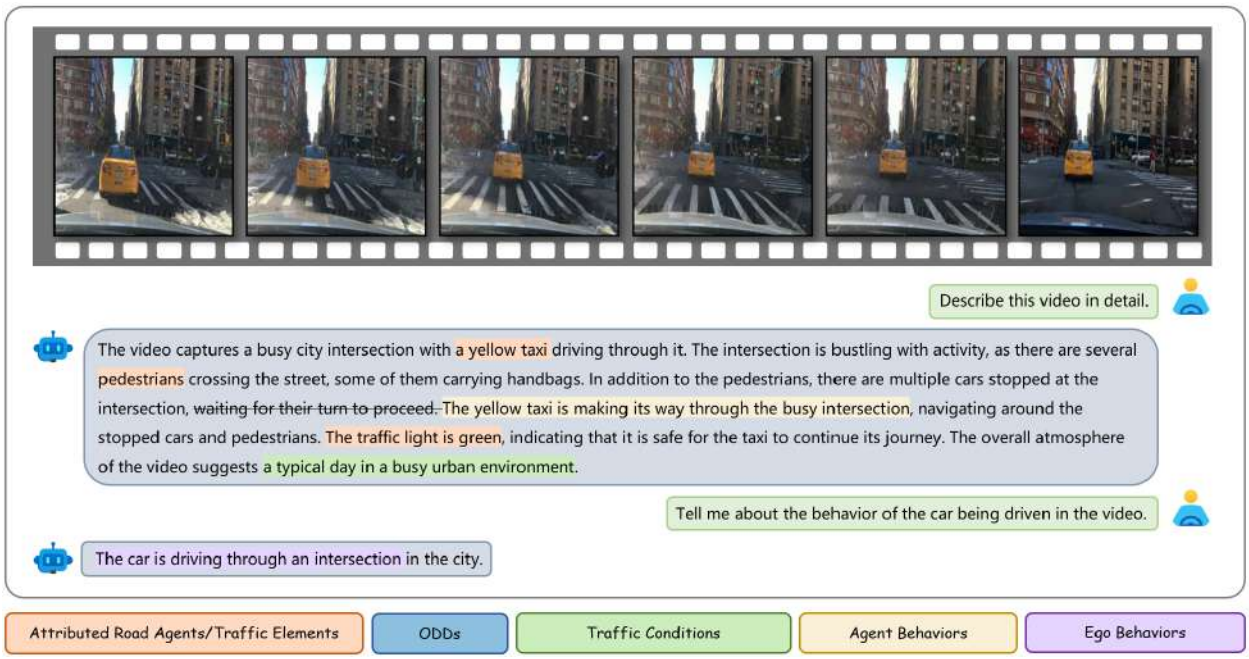

Figure 9: An example showcasing `Dolphins`'s capability in **scenario understanding** and **behavior understanding** § 5.1.1. The video shows an ego car following a taxi and going through an intersection. ~~Words~~ means hallucination.

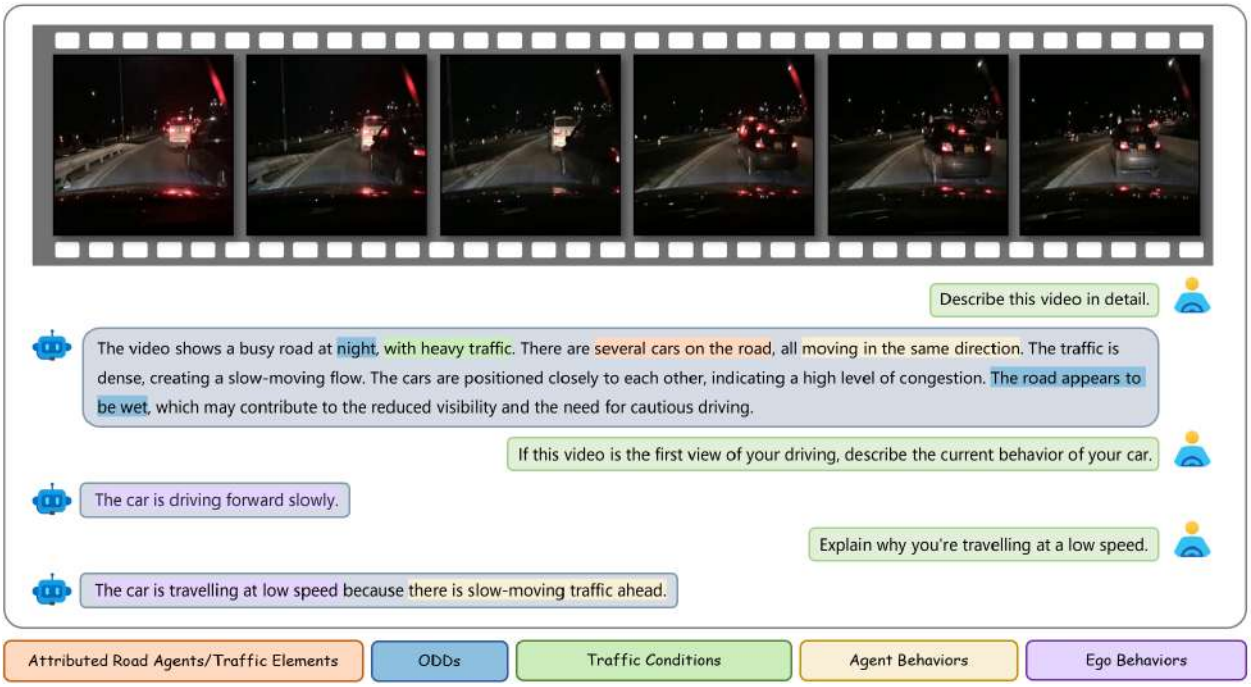

Figure 10: An example showcasing `Dolphins`'s capability in **scenario understanding** and **behavior understanding** § 5.1.1. The video shows an ego car driving slowly on a busy road at night. `Dolphins` can identify the ego car traveling at a slow speed and infer that the reason is that the speed of the vehicle ahead is restricting the ego car's speed.

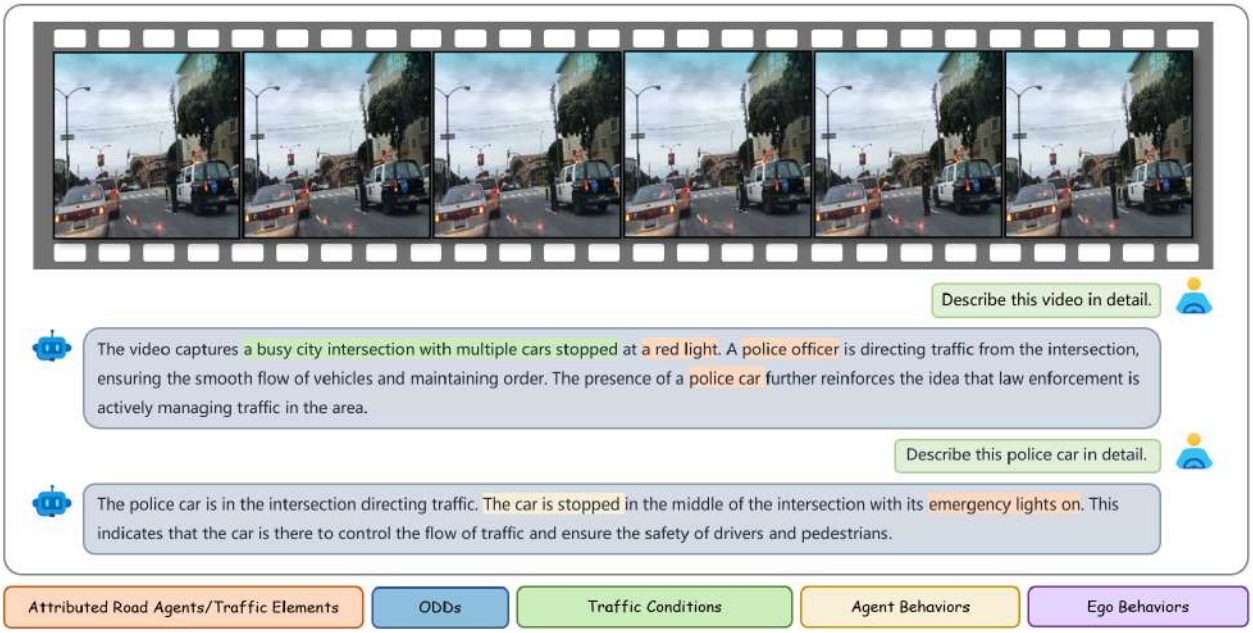

Figure 11: An example showcasing `Dolphins`'s capability in **scenario understanding** § 5.1.1. The video shows an ego car stopped at a busy interaction with a police car next to it. `Dolphins` can identify the police officer and police car with its emergency lights on.

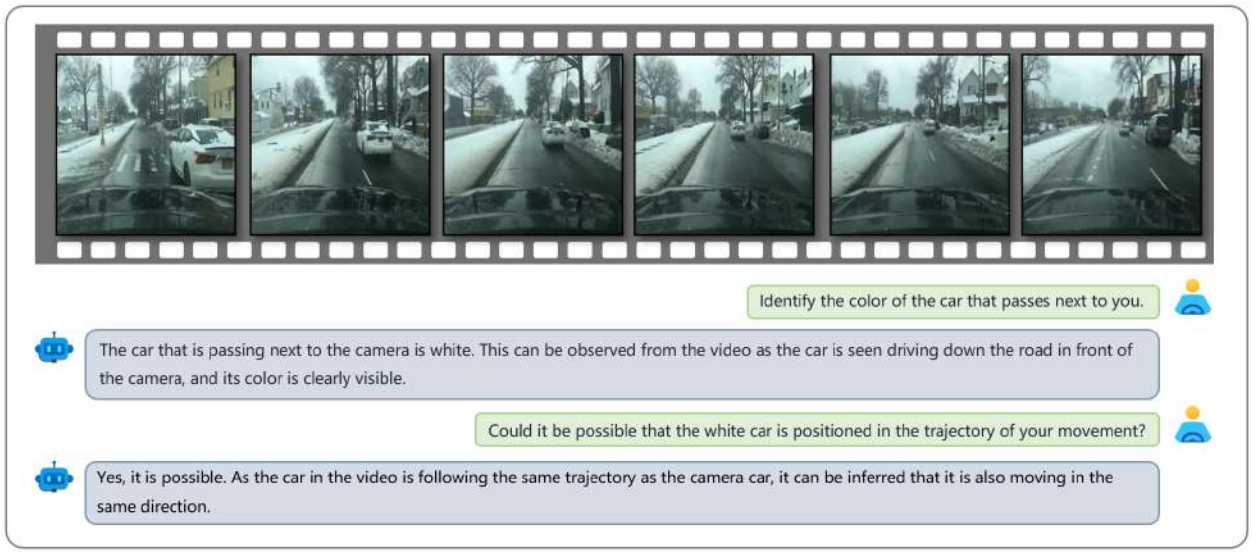

Figure 12: An example showcasing `Dolphins`'s capability in **prediction** § 5.1.2. The video features an ego car driving on a snowy street. `Dolphins` can predict the trajectory of the white car passing by it. Since both the white car and the ego car are traveling in the same direction on a one-way road, the trajectories of both cars temporarily overlap.

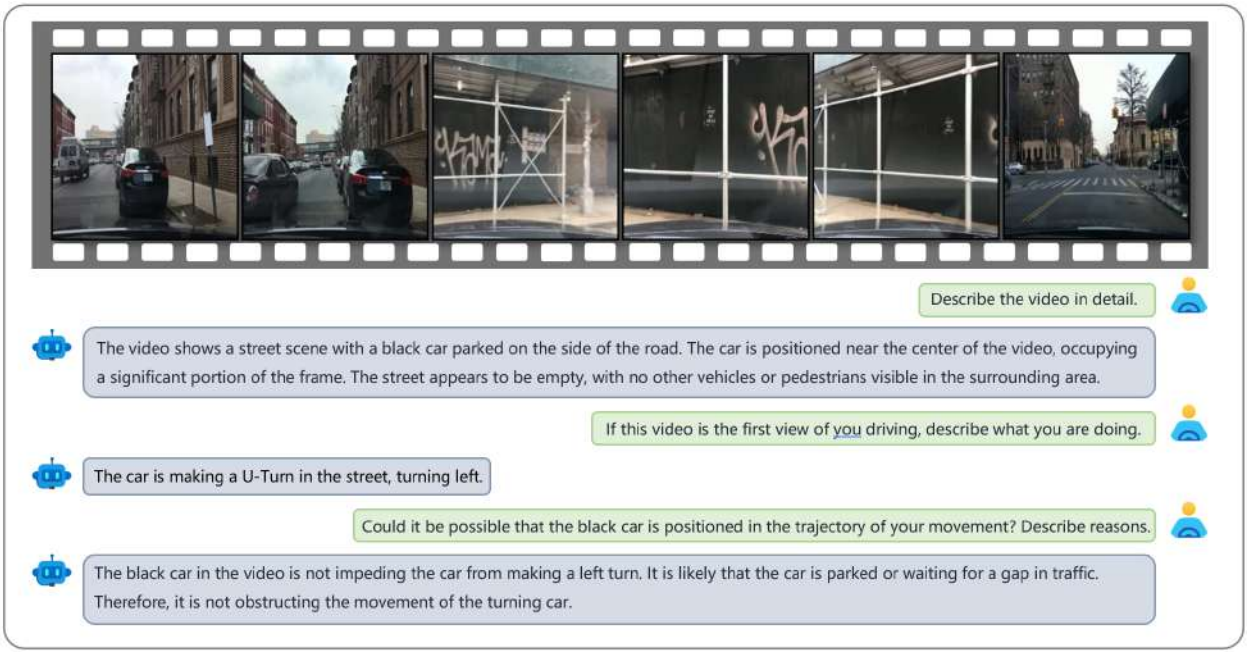

Figure 13: An example showcasing `Dolphins`'s capability in **prediction** § 5.1.2. The video shows an ego car making a U-turn. `Dolphins` can predict the trajectory of the black car parked in front of it. Since the black car in front is currently parked on the side of the road, Our model predicts that the car will remain there in the future and will not be positioned in the ego car's trajectory of turning.

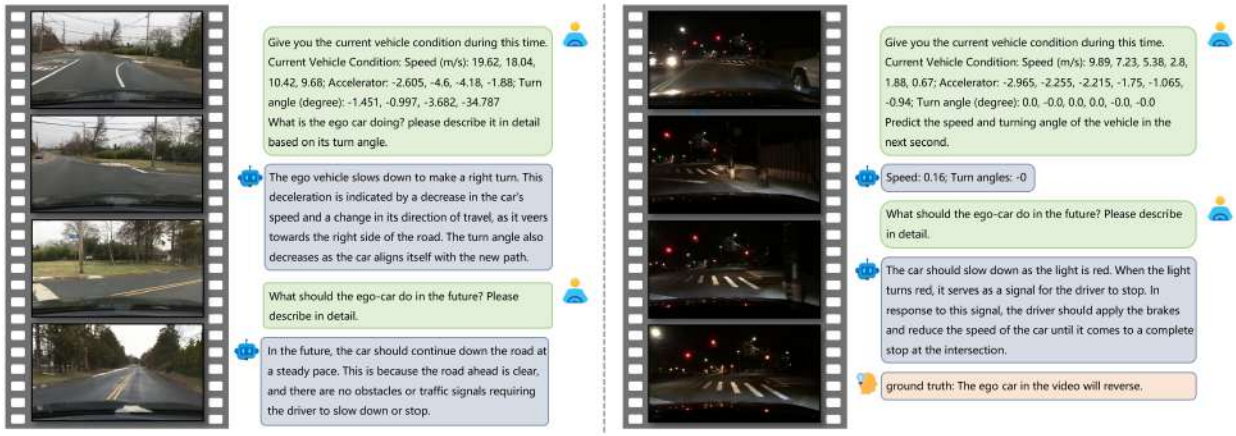

Figure 14: An example showcasing `Dolphins`'s capability in **planning** § 5.1.2 with control signals. `Dolphins` can also understand control signals and employ them to predict the speed and turn angle of the ego car next second. Additionally, We can enable `Dolphins` to formulate plans for the ego car, thereby assisting the driver during driving. However, due to the lack of sufficient information, such as the map, driving destination, etc., the plan is still limited to a brief period in the future.

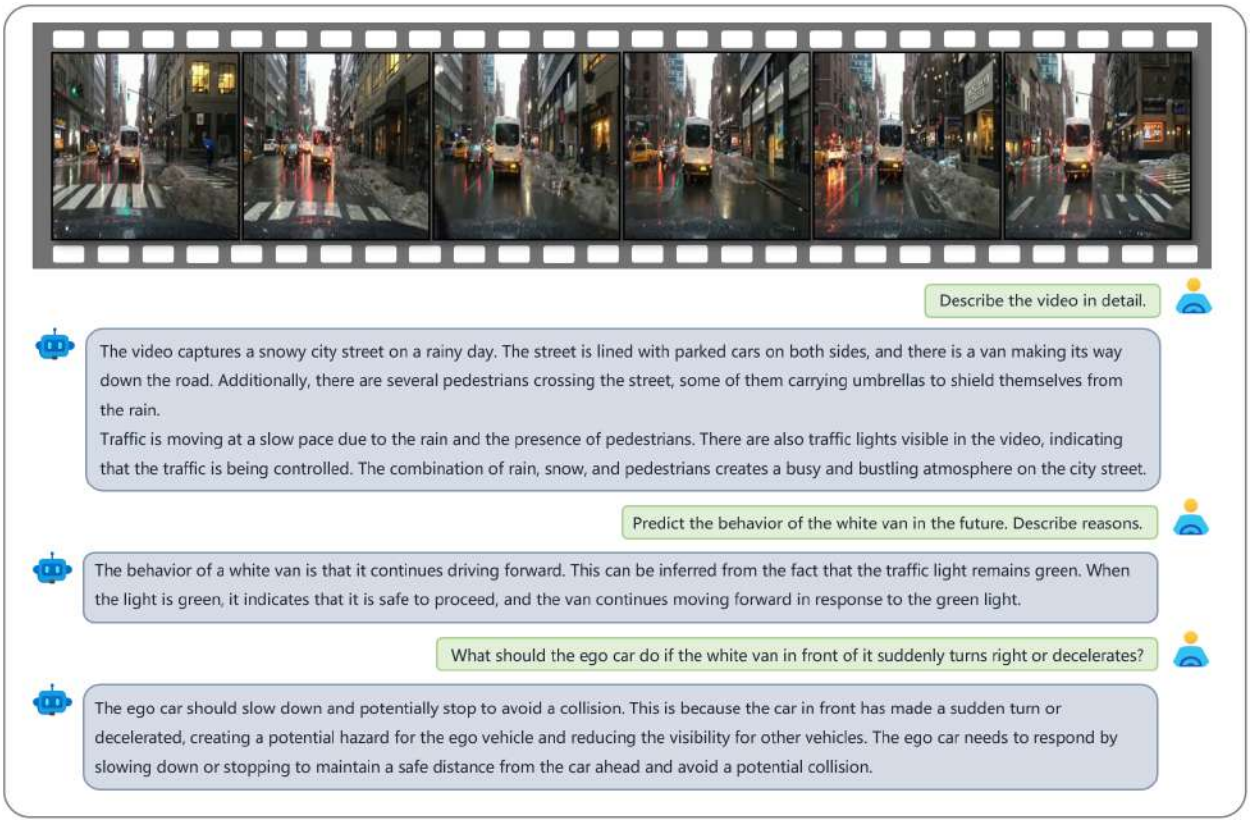

Figure 15: An example showcasing `Dolphins`'s capability in **prediction** and **contingency planning** § 5.1.2. The video features an ego car driving on a snowy city street on a rainy day. `Dolphins` can predict the future behavior of the white van in front of it while making reasonable contingency plans for the ego car in case the white decelerates or turns right.

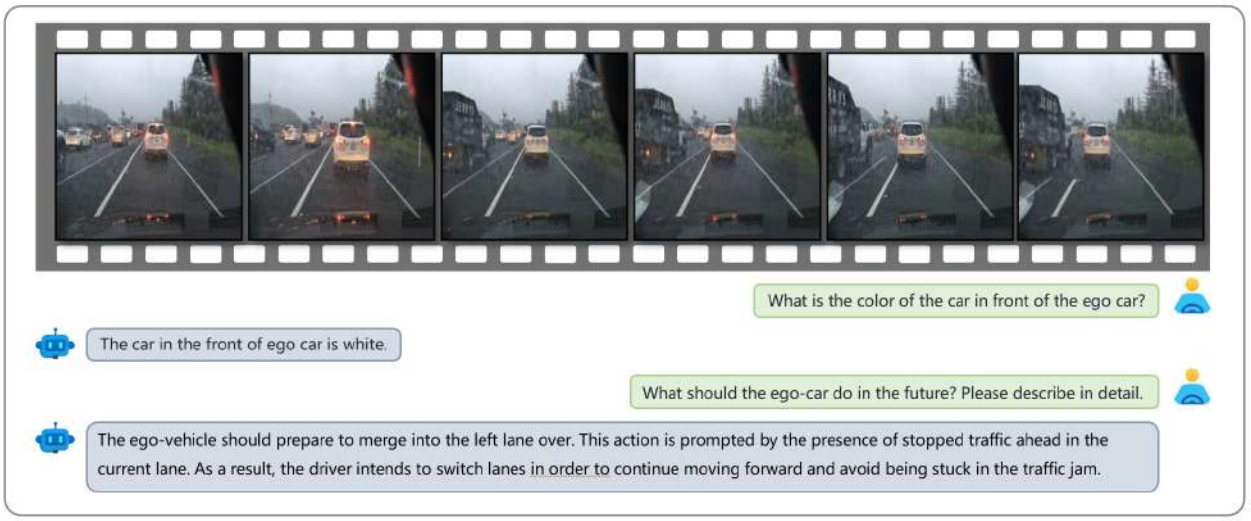

Figure 16: An example showcasing `Dolphins`'s capability in **planning** § 5.1.2. The video features an ego car driving on a highway and a white car is parked in front of it. `Dolphins`, by assessing that the white car in front has come to a stop, plans for the future behavior of the ego car, which should involve changing lanes to the left to avoid a collision with the stationary white car.

Figure 17: An example showcasing `Dolphins`'s capability in **planning** § 5.1.2. The video features an ego car driving on a highway on a rainy day. `Dolphins` is asked about the information regarding the black car to the right of the ego car, and it accurately determined that the black car's lane change would not affect the ego car's trajectory because they are separated by two lanes.

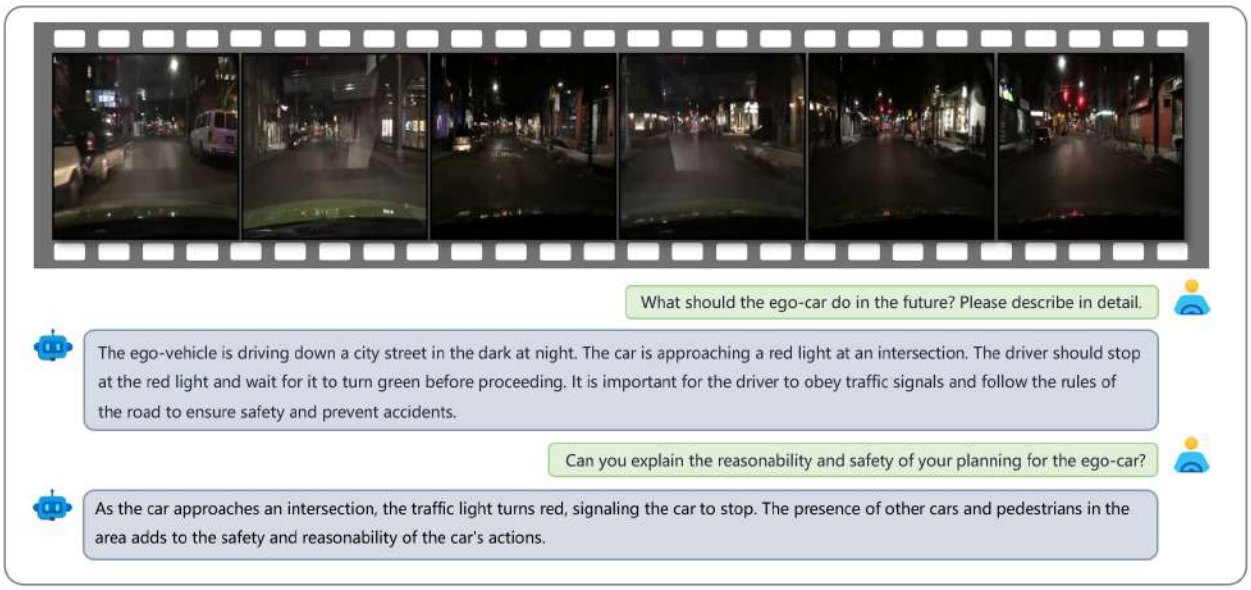

Figure 18: An example showcasing `Dolphins`'s capability in **planning** § 5.1.2. The video features an ego car driving on a dark city street. `Dolphins` can recognize that the ego car is approaching an intersection with a red traffic light, so it plans the future behavior for the ego car, which should be to come to a stop and wait for the traffic light to turn green to pass through the intersection safely.

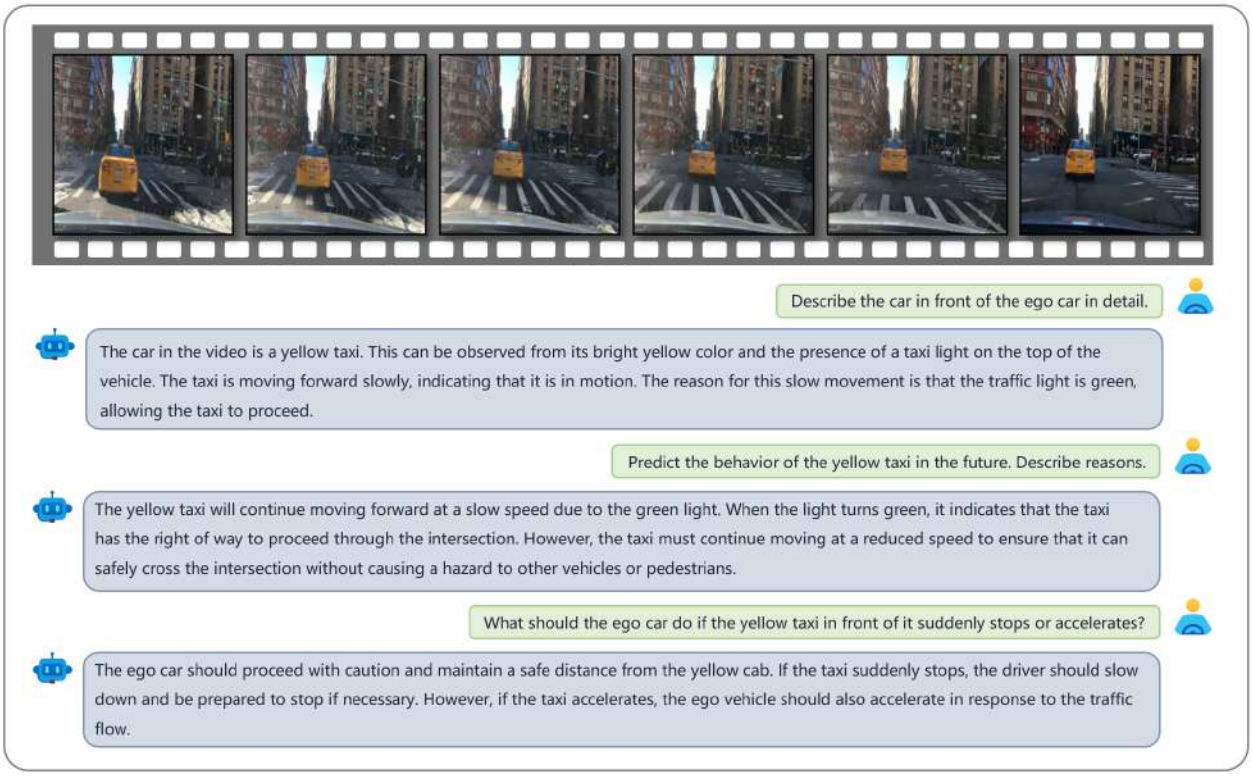

Figure 19: An example showcasing `Dolphins`'s capability in **prediction** and **contingency planning** § 5.1.2. The video shows an ego car following a taxi and going through an intersection. On one hand, `Dolphins` can predict the future behavior of the yellow taxi for a certain period. On the other hand, `Dolphins` can make reasonable contingency plans for the ego car in case the yellow taxi in front suddenly accelerates or comes to a stop.

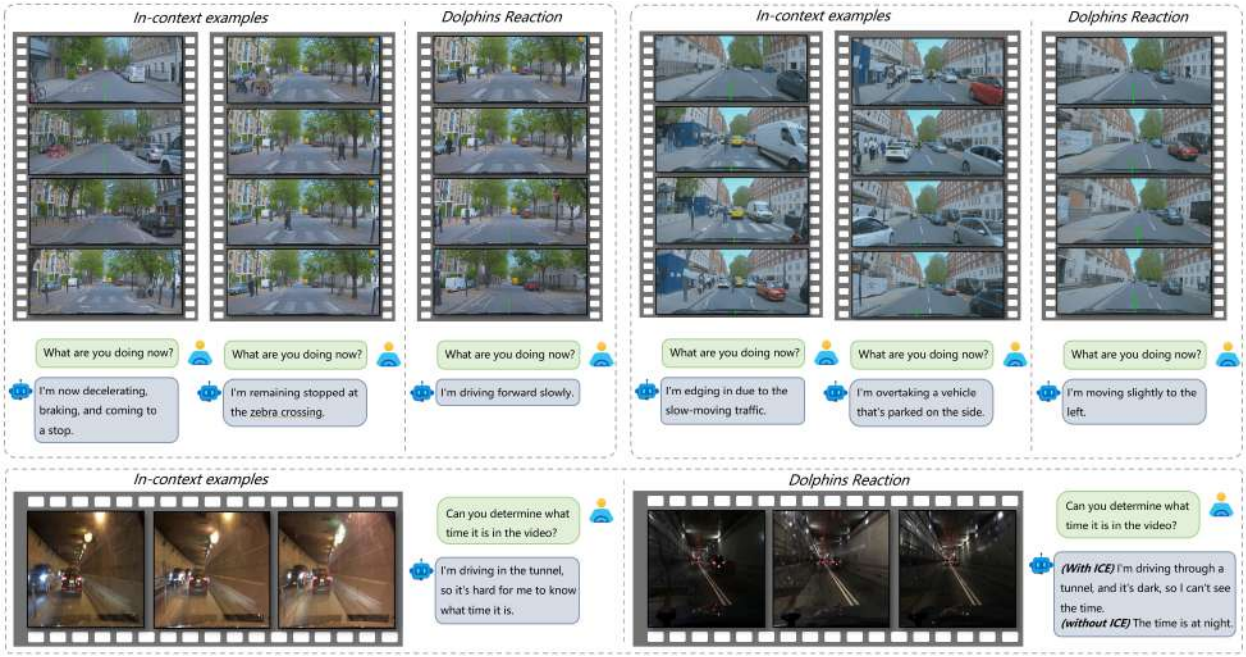

Figure 20: Three examples show our model enables rapid adaptation to unseen instructions through **in-context learning** § 5.2.1. In the first two examples, `Dolphins` learns to play the role of the driver through in-context examples and can accurately describe its behavior, despite not having been trained on such instructions. The third example shows that `Dolphins` can learn common sense from in-context examples, such as not being able to judge the current time based on the light when inside a tunnel.

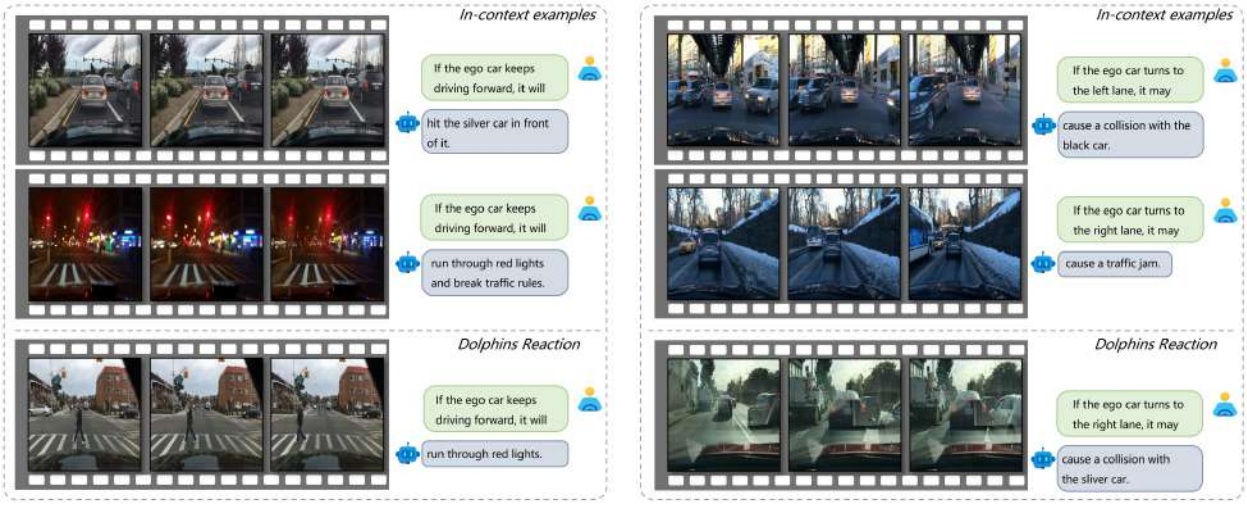

Figure 21: An example showcasing `Dolphins`'s capability in **prediction** through **in-context learning** § 5.2.1. `Dolphins` can reason for future events that have not yet happened. Through in-context examples, our model can learn to answer "What If"-style questions, to imagine the behavior of the ego car in the future by language.

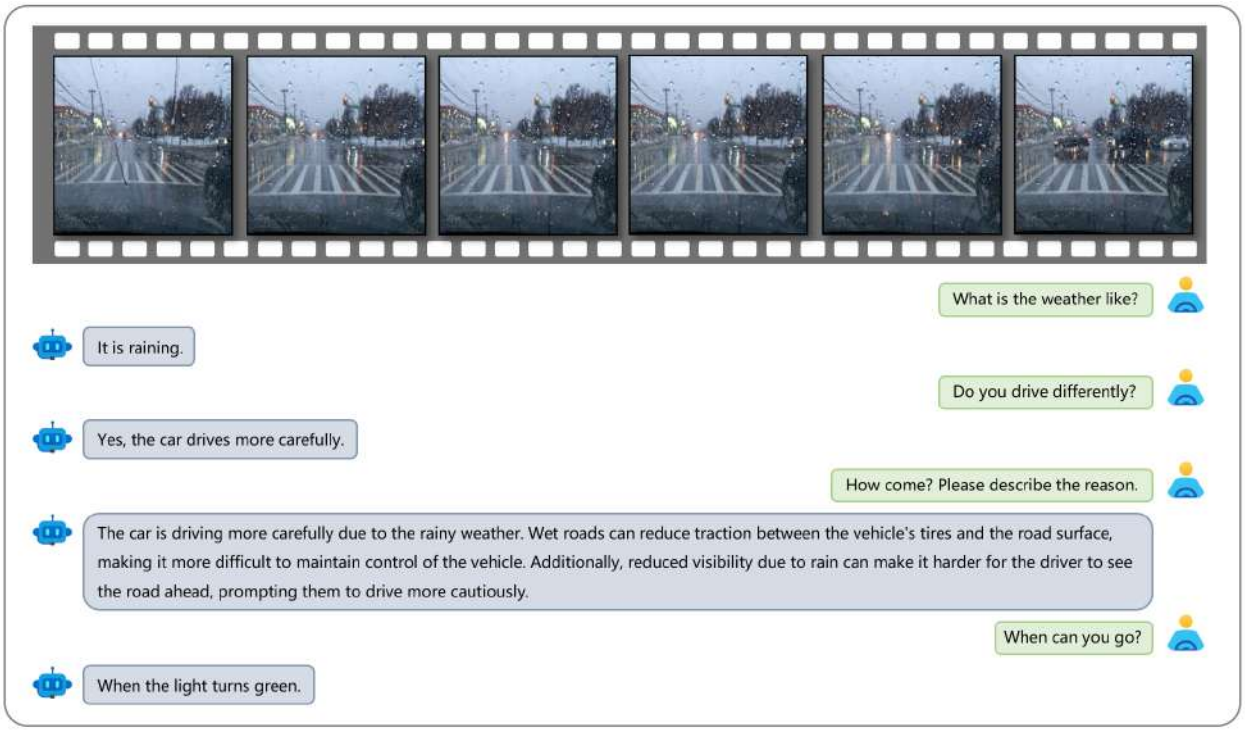

Figure 22: An example showcasing `Dolphins`'s capability in **interactive conversation** § 5.2.2. The video features an ego car stopping at an intersection on a rainy day, where `Dolphins` can answer various questions about the weather, the traffic lights, and their impact on driving.

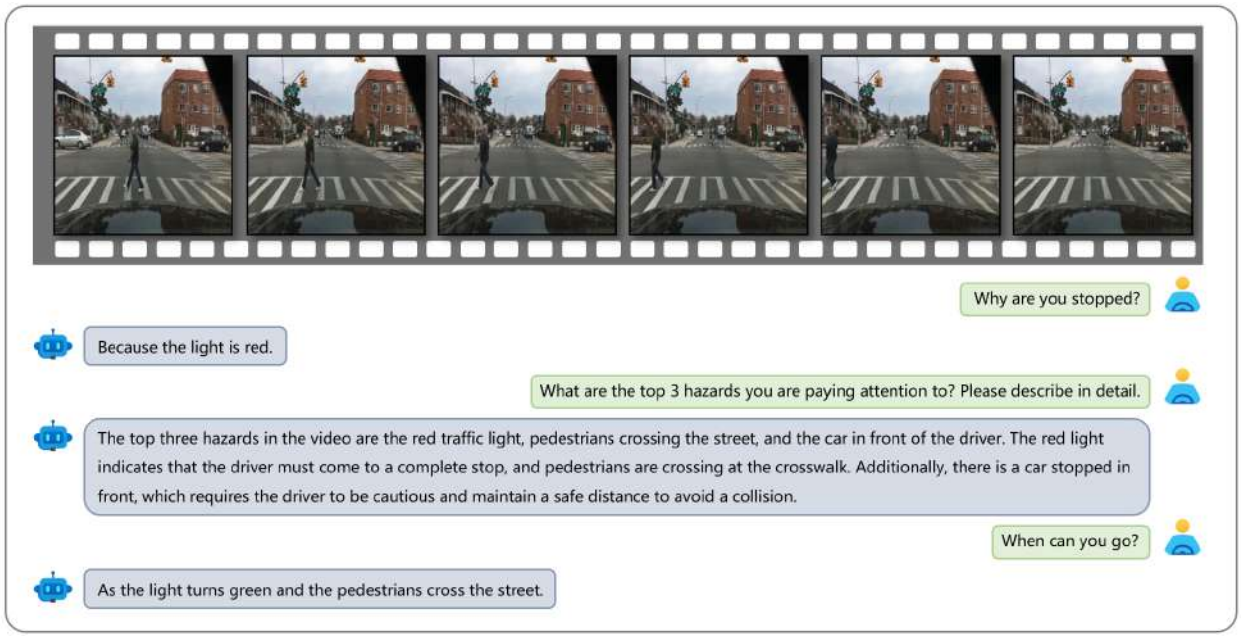

Figure 23: An example showcasing `Dolphins`'s capability in **interactive conversation** § 5.2.2. This video shows an ego car stopped at an intersection waiting for a red light and a pedestrian crossing a zebra crossing. We can ask `Dolphins` some questions about what it's paying attention to at this intersection.

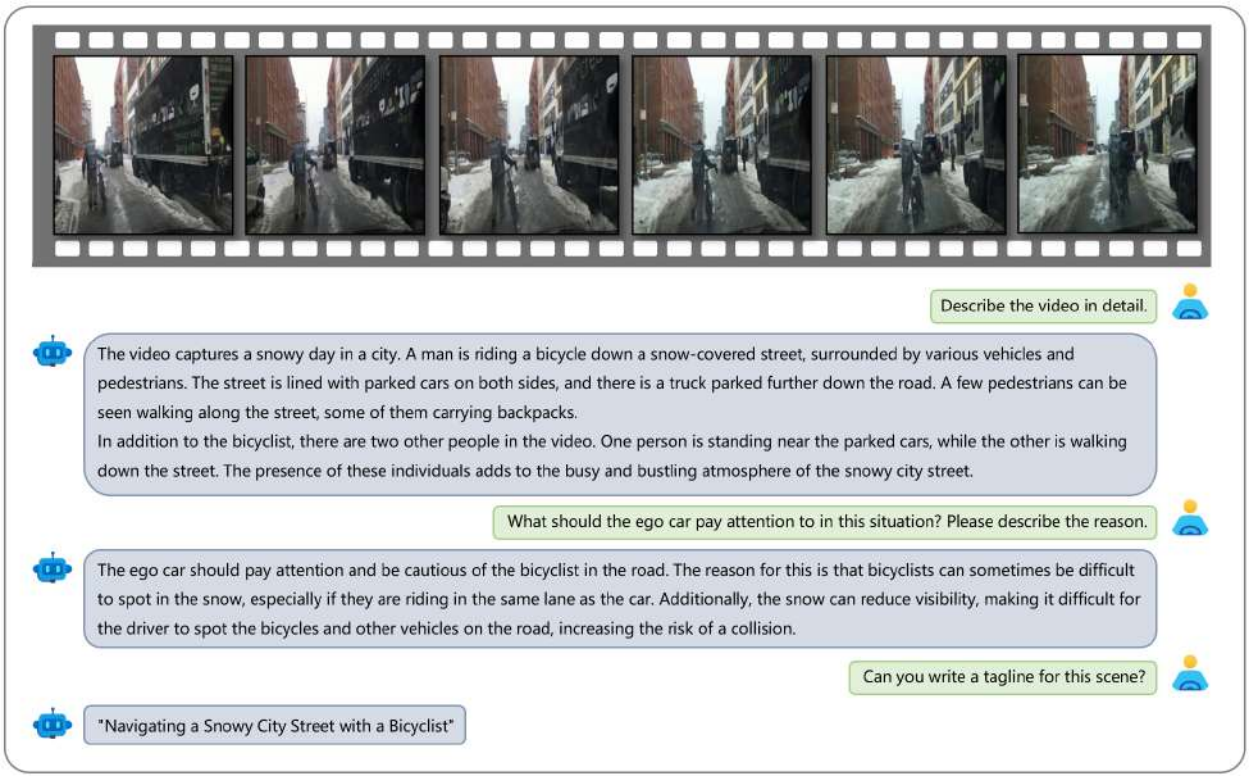

Figure 24: An example showcasing `Dolphins`'s capability in **interactive conversation** § 5.2.2. This video shows an ego car driving behind a bicyclist. Through conversation with `Dolphins`, we can understand how to drive safely in the presense of cyclists. And our model can also create a tagline for this scene.

Figure 25: An example showcasing `Dolphins`'s capability in **reflection** § 5.2.3. The video features an ego car driving through an interaction on a city street. The black vehicle in front of the ego car appears to want to turn right as its right taillight is flashing. Initially, `Dolphins` predicts that the black car will continue to drive forward. However, after being provided with information about the 'slashing rear lights indicate a turn or lane change.' traffic rule, `Dolphins` reflects and revises the wrong prediction.

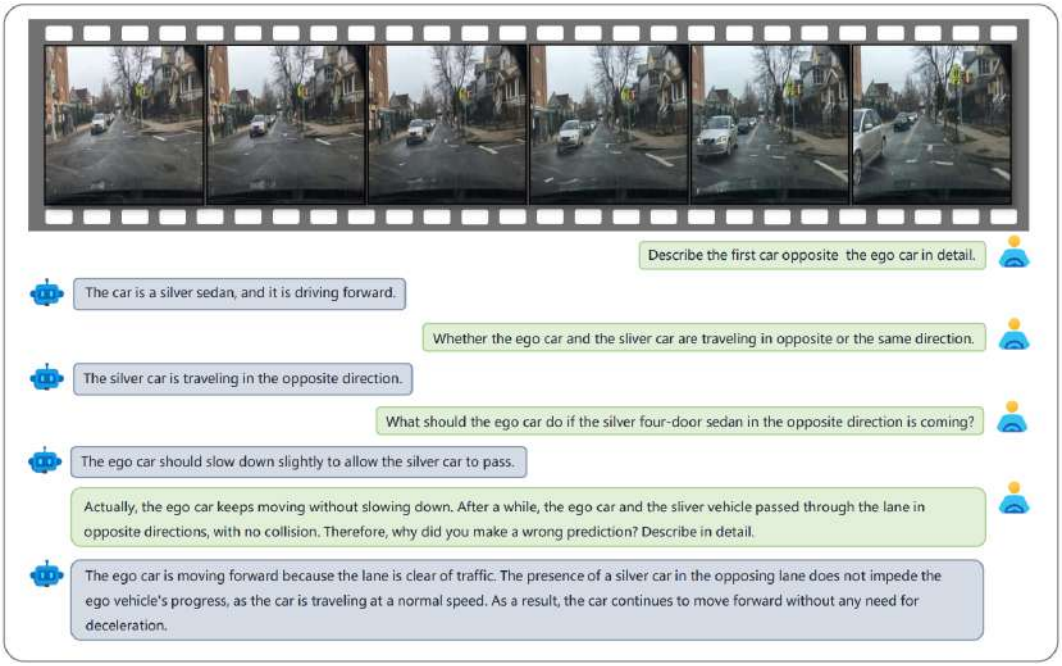

Figure 26: An example showcasing `Dolphins`'s capability in **reflection** § 5.2.3. The video shows an ego car driving through an interaction with a white car coming from the opposite direction. Initially, `Dolphins` plans for the ego car to slow down and let the white car proceed. However, after being provided with information about the human driver decision, `Dolphins` reflects and revises the driving plan.

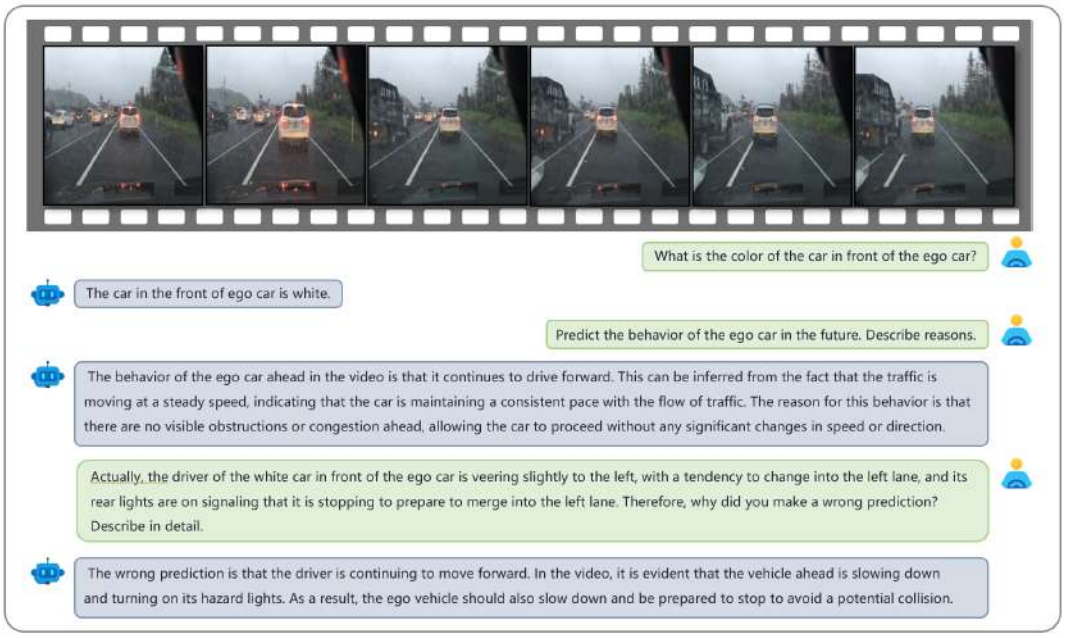

Figure 27: An example showcasing `Dolphins`'s capability in **reflection** § 5.2.3. The video shows an ego car driving on a highway. Initially, `Dolphins` predicts that the white car in front of the ego car will continue to drive forward. However, after being provided with information about the human driver decision, `Dolphins` revises the wrong prediction and realizes that the ego car should have also stopped to avoid the white car that is about to stop and merge into the left lane.

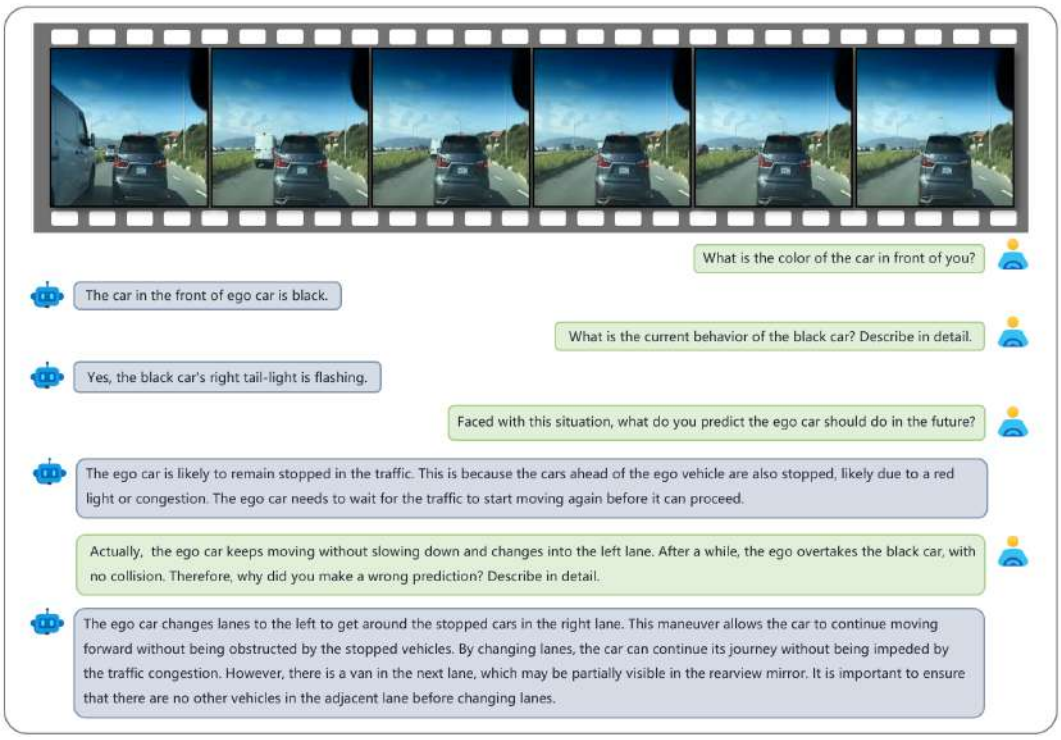

Figure 28: An example showcasing `Dolphins`'s capability in **reflection** § 5.2.3. The video shows an ego car stopping behind a black car. Initially, `Dolphins` plans for the ego car to remain stopped and wait for the traffic. However, after being provided with information about the human driver decision (driving more aggressively), `Dolphins` reflects and plans for the ego car to change lanes to the left to continue moving forward when careful attention is paid to cars traveling the left lane, such as a white van.

References

[1] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. Science, 349(6245):255–260, 2015.

[2] Mohsen Soori, Behrooz Arezoo, and Roza Dastres. Artificial intelligence, machine learning and deep learning in advanced robotics, a review. Cognitive Robotics, 2023.

[3] Gijs Mom. The evolution of automotive technology: a handbook. SAE International, 2023.

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

[5] Ardi Tampuu, Tambet Matiisen, Maksym Semikin, Dmytro Fishman, and Naveed Muhammad. A survey of end-to-end driving: Architectures and training methods. IEEE Transactions on Neural Networks and Learning Systems, 33(4):1364–1384, 2020.

[6] Daniel Coelho and Miguel Oliveira. A review of end-to-end autonomous driving in urban environments. IEEE Access, 10:75296–75311, 2022.

[7] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. arXiv preprint arXiv:2306.16927, 2023.

[8] Liangkai Liu, Sidi Lu, Ren Zhong, Baofu Wu, Yongtao Yao, Qingyang Zhang, and Weisong Shi. Computing systems for autonomous driving: State of the art and challenges. IEEE Internet of Things Journal, 8(8):6469–6486, 2020.

[9] Ashesh Jain, Luca Del Pero, Hugo Grimmett, and Peter Ondruska. Autonomy 2.0: Why is self-driving always 5 years away? arXiv preprint arXiv:2107.08142, 2021.

[10] Kelvin Wong, Yanlei Gu, and Shunsuke Kamijo. Mapping for autonomous driving: Opportunities and challenges. IEEE Intelligent Transportation Systems Magazine, 13(1):91–106, 2020.

[11] Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9329–9338, 2019.

[12] Lucas de Paula Veronese, Fernando Auat-Cheein, Filipe Mutz, Thiago Oliveira-Santos, José E Guivant, Edilson De Aguiar, Claudine Badue, and Alberto Ferreira De Souza. Evaluating the limits of a lidar for an autonomous driving localization. IEEE Transactions on Intelligent Transportation Systems, 22(3):1449–1458, 2020.

[13] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al. Towards fully autonomous driving: Systems and algorithms. In 2011 IEEE intelligent vehicles symposium (IV), pages 163–168. IEEE, 2011.

[14] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. A survey of autonomous driving: Common practices and emerging technologies. IEEE access, 8:58443–58469, 2020.

[15] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. ArXiv, abs/2304.08485, 2023.

[16] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971, 2023.

[17] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890, 2023.

[18] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. ArXiv, abs/2308.01390, 2023.

[19] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 6904–6913, 2017.

[20] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pages 3195–3204, 2019.

[21] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6700–6709, 2019.

[22] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In Proceedings of the IEEE international conference on computer vision, pages 1965–1973, 2017.

[23] Bo Zhao, Boya Wu, and Tiejun Huang. Svit: Scaling up visual instruction tuning. arXiv preprint arXiv:2307.04087, 2023.

[24] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John F. Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In European Conference on Computer Vision, 2018.

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

[26] Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. ArXiv, abs/2310.01415, 2023.

[27] Ye Jin, Xiaoxi Shen, Huiling Peng, Xiaoan Liu, Jingli Qin, Jiayang Li, Jintao Xie, Peizhong Gao, Guyue Zhou, and Jiangtao Gong. Surrealdriver: Designing generative driver agent simulation framework in urban contexts based on large language model. arXiv preprint arXiv:2309.13193, 2023.

[28] DriveLM Contributors. Drivelm: Drive on language. https://github.com/OpenDriveLab/DriveLM, 2023.

[29] Dongming Wu, Wencheng Han, Tiancai Wang, Ying-Hao Liu, Xiangyu Zhang, and Jianbing Shen. Language prompt for autonomous driving. ArXiv, abs/2309.04379, 2023.

[30] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. arXiv preprint arXiv:1903.11027, 2019.

[31] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. ArXiv, abs/2302.13971, 2023.

[32] Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. ArXiv, abs/2307.09288, 2023.

[33] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[34] MosaicML NLP Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. Accessed: 2023-05-05.

[35] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. ArXiv, abs/2204.14198, 2022.

[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. ArXiv, abs/2301.12597, 2023.

[37] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. ArXiv, abs/2304.10592, 2023.

[38] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. ArXiv, abs/2305.03726, 2023.

[39] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. ArXiv, abs/2305.06500, 2023.

[40] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yi Zhou, Junyan Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qiang Qi, Ji Zhang, and Feiyan Huang. mplug-owl: Modularization empowers large language models with multimodality. ArXiv, abs/2304.14178, 2023.

[41] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355, 2023.

[42] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858, 2023.

[43] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424, 2023.

[44] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Minghui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. arXiv preprint arXiv:2306.07207, 2023.

[45] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378, 2023.

[46] Yao Mu, Qinglong Zhang, Mengkang Hu, Wenhai Wang, Mingyu Ding, Jun Jin, Bin Wang, Jifeng Dai, Yu Qiao, and Ping Luo. Embodiedgpt: Vision-language pre-training via embodied chain of thought. arXiv preprint arXiv:2305.15021, 2023.

[47] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818, 2023.

[48] Jingkang Yang, Yuhao Dong, Shuai Liu, Bo Li, Ziyue Wang, Chencheng Jiang, Haoran Tan, Jiamu Kang, Yuanhan Zhang, Kaiyang Zhou, et al. Octopus: Embodied vision-language programmer from environmental feedback. arXiv preprint arXiv:2310.08588, 2023.

[49] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. arXiv preprint arXiv:2307.12981, 2023.

[50] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. arXiv preprint arXiv:2308.16911, 2023.

[51] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. arXiv preprint arXiv:2304.08247, 2023.

[52] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: a multimodal medical few-shot learner. arXiv preprint arXiv:2307.15189, 2023.

[53] Ziqiang Zheng, Jipeng Zhang, Tuan-Anh Vu, Shizhe Diao, Yue Him Wong Tim, and Sai-Kit Yeung. Marinegpt: Unlocking secrets of ocean to the public. arXiv preprint arXiv:2310.13596, 2023.

[54] Xinpeng Ding, Jianhua Han, Hang Xu, Wei Zhang, and Xiaomeng Li. Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving. arXiv preprint arXiv:2309.05186, 2023.

[55] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kenneth K.Y. Wong, Zhenguo Li, and Hengshuang Zhao. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. ArXiv, abs/2310.01412, 2023.

[56] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, C. Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. ArXiv, abs/2306.05425, 2023.

[57] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. ArXiv, abs/2110.15943, 2021.

[58] Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. arXiv preprint arXiv:2307.03601, 2023.

[59] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195, 2023.

[60] Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. arXiv preprint arXiv:2112.08633, 2021.

[61] Openai chat. https://chat.openai.com. Accessed: 2023-10-20.

[62] Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. Meta-learning via language model in-context tuning. ArXiv, abs/2110.07814, 2021.

[63] Srinivas Iyer, Xiaojuan Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Veselin Stoyanov. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. ArXiv, abs/2212.12017, 2022.

[64] S. Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. In International Conference on Machine Learning, 2023.

[65] Julian Coda-Forno, Marcel Binz, Zeynep Akata, Matthew M. Botvinick, Jane X. Wang, and Eric Schulz. Meta-in-context learning in large language models. ArXiv, abs/2305.12907, 2023.

[66] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. ArXiv, abs/2210.08402, 2022.

[67] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. ArXiv, abs/2304.06939, 2023.

[68] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In Advances in Neural Information Processing Systems, 2022.

[69] Anonymous. Understanding multimodal instruction format for in-context learning. In Submitted to The Twelfth International Conference on Learning Representations, 2023. under review.

[70] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748–8763. PMLR, 2021.

[71] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021.

[72] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. ArXiv, abs/2306.02858, 2023.

[73] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wen Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. ArXiv, abs/2305.06355, 2023.

[74] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. ArXiv, abs/2306.05424, 2023.

[75] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Ming-Hui Qiu, Pengcheng Lu, Tao Wang, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. ArXiv, abs/2306.07207, 2023.

[76] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 3505–3506, 2020.

[77] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations, 2018.

[78] Tinychat: Large language model on the edge. https://hanlab.mit.edu/blog/tinychat. Accessed: 2023-10-20.

[79] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014.

[80] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision, 123:32–73, 2017.

A Data

Image instruction-following dataset enriched with GCoT response. The majority of vision-language tasks can be generally viewed as Visual Question Answering (VQA) tasks, requiring the model to provide answers to queries related to the image. Therefore, we collect 4 VQA datasets to generate GCoT response by ChatGPT, including VQAv2 [19], OK-VQA [20], GQA [21], and TDIUC [22]. Except for GQA, the image source for these tasks is MSCOCO [79], which contains many images, but each image has fewer annotations of caption and object. This may result in the object from the question not having corresponding position information during step (2), making it difficult for ChatGPT to provide an accurate reasoning process. Therefore, we use the Visual Genome dataset [80] as a supplement, as it has richer annotations and intersects with MSCOCO. The GQA task provides detailed object annotations but lacks captions, which presents a challenge to ChatGPT in comprehending the overall content of the image. So we organize the objects, attributes, and their relationships in the annotations into sentences, which are used to describe the relationships between two objects in the image in place of captions. After preparation, we prompt ChatGPT to follow the aforementioned three steps to generate GCoT templates step by step. The prompts can be found in Table 2. In addition, we also include LLaVA-instruct-80k [15] and SVIT [23] datasets to enhance the model's instruction-following capability. In summary, in the first stage, `Dolphins` is trained on an image instruction-following dataset comprising approximately 10.7k examples. Within this dataset, there are 9,645 VQA examples accompanied by GCoT responses, which are generated by ChatGPT.

Video instruction-following dataset based on BDD-X. To transition the model's powerful scene understanding and reasoning ability, which has been fine-tuned on the image instruction-following datasets, to the driving video domain, we construct an autonomous driving-related instruction-following dataset based on BDD-X, and at the same time retrieve in-context examples to generate few-shot templates for training to enhance model's in-context learning ability by retrieve method.

B Prompts

The prompts used to instruct ChatGPT to generate the grounded CoT process with three thinking steps for VQA tasks are shown in Table 2.

System Message

Give you some captions, each describing the image you are observing and specific object locations within the image are given, along with detailed coordinates. These coordinates are in the form of bounding boxes, represented as (x1, y1, x2, y2) with floating numbers ranging from 0 to 1. These values correspond to the top left x, top left y, bottom right x, and bottom right y. Besides, you will receive a question about the image and the reference answer to this question.

I want you to create a paragraph to explain how to obtain the reference answer for the question. To create the paragraph, follow these steps:

1. Create a short sentence to describe the image by using captions.
2. Find the question object contained in the question. If the question object exists in the provided objects, create a sentence describing the object and its bounding box; Instead, find the most relevant one with the question object from the provided objects and create a sentence describing the object and its bounding box.
3. If the answer to the question requires reasoning, create a sentence to explain the question. For instance, the question is "When was this piece of sporting equipment invented?" and if in step 3 you have determined that the "sports equipment" is a surfboard, based on your background knowledge, the sentence you need to create is: "In 1926 an American surfer named Tom Blake (1902 - 1994) invented the very first, hollow surfboard."; If the answer to the question does not require reasoning and can be derived directly by looking at the image, for example, the question "What color of the bear is?", then in this step, you do not need to create any sentences to explain it.

Note: You need to determine the position between the objects based on their bounding box. For example object A: [x1, y1, x2, y2], object B: [a1, b1, a2, b2]. If $x2 > a2$, it means that object A is to the right of object B, and vice versa. If $y2 > b2$, it means that object A is on top of object B, and vice versa.

User

Captions:
Man in all black swimsuit walking down a beach with his surfboard.
A man in a wetsuit carrying a surfboard to the water.
A person with a surfboard walking on a beach.
A person with a surfboard walks to the water.
A man carrying a surfboard across a sandy beach.

Objects:
bird: [0.095, 0.797, 0.355, 0.849], surfboard: [0.388, 0.418, 1.254, 0.977], person: [0.431, 0.222, 0.941, 1.362]

Question: When was this piece of sporting equipment invented?
Reference Answer: 1926

Output

1. The picture shows a man carrying a surfboard across a sandy beach.
2. So the sporting equipment in question should refer to the surfboard.
3. In 1926 an American surfer named Tom Blake (1902 - 1994) invented the very first, hollow surfboard.

Table 2: One example to prompt ChatGPT for generating GCoT to enhance the VLM's capabilities of fine-grained multimodal understanding and reasoning, which are considered the important proficiency in ADs.