Dolphins: Multimodal Language Model for Driving

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

Chaowei Xiao[1,2]
{g19myz}@gmail.com

[1] University of Wisconsin–Madison [2] NVIDIA
[3] University of Michigan-Ann Arbor [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 in the general domain. 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. The anonymous demo is available at 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 [25], robotics [50], and automotive engineering [44]. The essential aspiration is to design autonomous vehicles (AVs) capable of maneuvering through complex real-world driving situations with human-like understanding and responsiveness.

However, current autonomous driving systems (ADS) [10,12,33,36,51] exhibit many limitations when compared with human drivers including: (1) **Holistic Understanding and Interpretation**: unlike human drivers could immediately deduce the potential danger and act accordingly to prevent any mishap even in

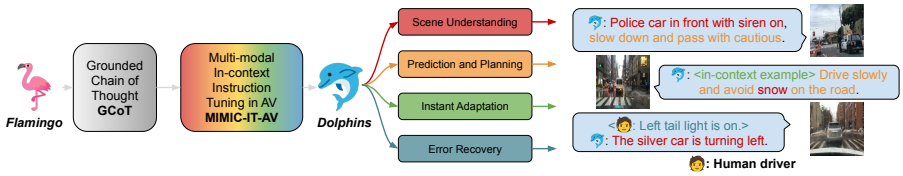

Fig. 1: Dolphins excels with abilities mirroring human drivers, from nuanced decision-making to fast adaptation with few-shot demonstrations (in-context learning) and error correction. Engineered for versatility, Dolphins also adeptly handle a wide range of driving tasks (e.g., perception, prediction, and planning).

the rare situation, 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 [22, 60]. (2) **Instant Learning and Adaptation**: unlike human drivers who can instantly learn and adapt to new scenarios, existing ADS requires extensive training with large amounts of data to handle new situations. (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.

Recent advancements in (multimodal) large language models (LLMs) [31, 34, 55] 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, where language data are often limited in both diversity and richness [13, 28].

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. To overcome the challenge of limited task diversity and text richness for language data in the AV context, we adapt **Dolphins** 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 Chain of Thought (GCoT) responses based on public VQA datasets [19, 21, 26, 43], visual instruction datasets [34, 69], and ChatGPT, to enable fine-grained reasoning capability in the OpenFlamingo model. Then, we utilize BDD-X [28] to establish our driving-related instruction dataset, focusing on four key AV tasks: behavior comprehension, control signal forecasting, behavior analysis, and in-depth conversation. We also employ the RICES (Retrieval-based In-Context Example Selection) approach [47] to choose in-context examples for each sample of the BDD-X training split. At last, we conduct in-context instruction tuning on our proposed dataset, grounding the general capabilities of the model to the AV

domain while enhancing its in-context learning ability for rapid adaptation to diverse new instructions.

Although being fine-tuned solely on a dataset comprising 32k QA-pairs across four driving-related tasks, our model still demonstrates its remarkable zero-shot and few-shot capabilities in diverse AV tasks. For example, **Dolphins** shows an advanced understanding of complex driving scenarios and human-like abilities such as instant adaptation, reflection, and reasoning. As a result, **Dolphins** showcases broad task applicability across perception, prediction, and planning, thanks to its comprehensive scenario understanding.

We also conduct comprehensive experiments across three benchmarks. On the test split of the BDD-X dataset, our model achieves the action of 223.6 CIDEr and the justification of 134.2 CIDEr, outperforming sota LVLMs [62,65] by 27% in the justification task. Furthermore, **Dolphins** exhibits remarkable generalization capabilities to diverse unseen autonomous driving instructions, encompassing perception, prediction, and planning. The results indicate that **Dolphins** surpasses existing video-based LVLMs on our proposed Scenario Understanding Benchmark and DriveLM by margins of 8.1 and 9.7 points, respectively. Additionally, by incorporating answers generated by **Dolphins** on unseen instructions back into the training dataset, we observe an incremental improvement of 1.1 points on DriveLM.

We summarize our contribution as four 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 conduct comprehensive evaluation on the effectiveness of the proposed method across three benchmarks including one built by us. The results demonstrate notable improvements in diverse tasks on several settings, including fine-tuning, zero-shot, and few-shots.
- We showcase the prominent capability of **Dolphins** spanning scene understanding and reasoning, instant adaptation, and error recovery, illustrated through numerous examples.

2 Related Work

Autonomous Driving with LLM. 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 [18] designs a new paradigm to mimic the process of human learning to drive based on LLMs while GPT-Driver [40] and Agent-Driver [41] leverages GPT-3.5 to assist autonomous driving in dependable

motion planning. In a parallel vein, SurrealDriver [24] 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 [13] and NuPrompt [61] introduce innovative driving tasks based on the NuScenes dataset [8]. 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 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 [11, 52, 54, 56] to enhance Large Vision Language Models (LVLMs), such as Flamingo [5] and BLIP-2 [32]. Recently, to unlock the capabilities of LVLMs to align with human preferences, LLaVA [34] and MiniGPT-4 [71] pioneers visual instruction tuning, with subsequent efforts like Otter [30] and InstructBLIP [15] following suit. Recently, LVLMs have been engineered to facilitate autonomous driving-related interaction and have already witnessed some related research efforts, which primarily encompassing key domains such as 2D/3D perception and high-level planning [16, 17, 42, 63], open-loop planning [13, 20, 53, 62, 65], and closed-loop planning [49, 59]. However, they all 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 [6] with strong in-context learning capabilities. Furthermore, we employ multimodal in-context instruction tuning [29, 70] 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.

3 Methodology

To equip LVLMs with 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 [45], 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 the creation of our

proposed dataset for autonomous driving in § 3.2 along with our devised tasks. Finally, we detail the multi-modal in-context instruction tuning for AV in § 3.3.

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 [9, 68]. Although HiLM-D [16] 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.

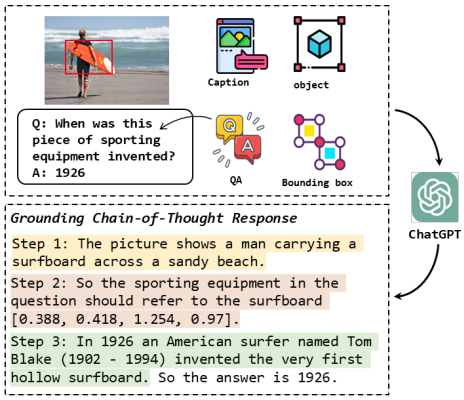

Fig. 2: The process of generating GCoT response for VQA tasks to enhance the fined-grained reasoning capability of VLMs. ChatGPT is prompted to generate GCoT step by step from text input.

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 B.

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.

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 [62] 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 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 [1] 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 [28]. 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 [62], we leverage the BDD-X dataset to develop our visual

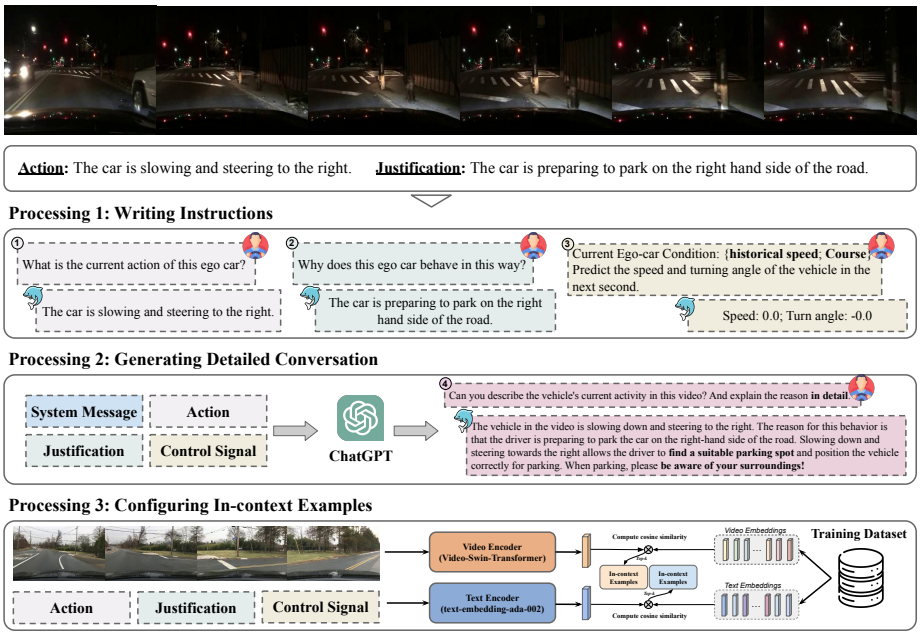

Fig. 3: Overview of our augmented dataset based on the BDD-X dataset. Firstly, we categorize the BDD-X origin dataset into three task types: action, justification, and control signal. For each task, we manually craft diverse instructions. Then, we introduce the "Detailed Conversation" task to unlock the latent potential of the foundation model that has been fine-tuned on the GCoT dataset consisting of image-instruction-response triplets for generating human-preferred detailed responses. Finally, compared with the previous datasets, we employ RICES (Retrieval-based In-Context Example Selection) [47] approach to choose in-context examples for each sample.

instruction-following dataset for autonomous driving, consisting of four distinct autonomous driving-related tasks and their corresponding instructions. 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 to achieve good results on the constructed dataset during the instruction tuning process.

3.3 In-context Instruction Tuning for Autonomous Driving

Due to limitations in the diversity of tasks and instructions in AV, the VLM trained on our augmented BDD-X dataset (Figure 3) exhibits a significant deficiency in its ability of zero-shot generalization to unseen tasks. Thus, we

leverage multi-modal in-context instruction tuning [30] to assist our model in the rapid adaptation to new instructions with just a handful of annotated examples in autonomous driving-related tasks.

In pursuit of the aforementioned objective, we employ OpenFlamingo [6] as our foundational VLM. OpenFlamingo, a reimplementation of Flamingo [5], is trained on the integration of image-text interleaved Lion-2B [48] and MMC4 [72] datasets to enhance its in-context learning capabilities. Our autonomous driving-related instruction dataset, as described in § 3.2, adopts a format comprising video-instruction-answer triplets. Consequently, we employ a retrieval approach to select in-context examples for each triplet. As shown in Figure 3, we utilize `Video-Swin-Transformer` [37], initialized with weights from ADAPT [23], and `text-embedding-ada-002` [4] 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. Finally, we utilize in-context examples retrieved by both text and image embedding similarity and constrain the provision of in-context examples to a maximum of $k = 3$ per triplet during the training stage.

4 Experiments and Results

In this section, we quantify **Dolphins**'s performance using a specially designed metric to assess its holistic understanding and reasoning skills. Additionally, we conduct an in-depth qualitative analysis to gauge its human-like capabilities, including adaptability in both zero- and few-shot settings.

4.1 Experimental Setups

Evaluation. We evaluate our model and baselines on three benchmarks: our proposed Scenario Understanding Benchmark, BDD-X [28], and DriveLM [13]. BDD-X is a dataset focused on driving-domain captions, where each video encompasses 1-5 behaviors along with the corresponding reasoning. DriveLM is an autonomous driving (AD) dataset that encompasses three distinct tasks, including perception, prediction, and planning, which are designed to introduce

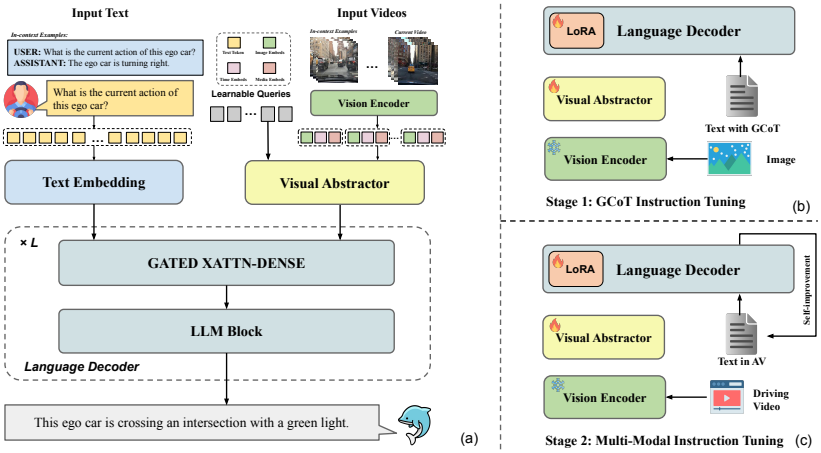

Fig. 4: Illustration of the proposed **Dolphins** and its training paradigm. (a) An overview of **Dolphins**. N videos are input into the vision encoder for extracting video features, with the first $N-1$ serving as in-context examples. The generated video features, augmented with media and frame position embeddings, are projected to visual embeddings by a visual abstractor. The LLM then processes visual and text embeddings to produce the final responses. (b) The training paradigm of **Dolphins** involves GCoT instruction tuning and multi-model in-context instruction tuning with updates exclusively made to the visual abstractor and LoRA parameters. Furthermore, the model can self-refine by training on its generated pseudo labels.

the LLM's or VLM's reasoning ability in autonomous driving (AD) to make decisions and ensure explainable planning.

Models and training details. In the first stage, **Dolphins-Image** is trained on the image-following instruction dataset with GCoT responses (Figure 4 (b)). Despite not being specifically trained on a video-based dataset, **Dolphins-Image** still possesses some capacity for video scenario understanding owing to the capacity of OpenFlamingo [6] to process multiple images as input. In the second stage, **Dolphins** is trained based on **Dolphins-Image** through in-context instruction tuning on the BDD-X dataset to transition the model's capabilities to the AV context (Figure 4 (c)). To enhance Dolphins, we introduce **Dolphins-Plus**, where the model self-refines by training on its generated pseudo labels from Scenario Understand Benchmark (Figure 4 (c)). This self-improvement [66] technique has gained traction in LLMs and we contend that it is a promising under-explored facet for LVLMs in the future. Details can be found in Appendix C.

Baselines. We conduct the experiments to compare with existing video-based VLMs, including OpenFlamingo [6], Otter-video [29], Video-LLaMA [67], Video-ChatGPT [39], and Valley [38]. In addition, we incorporate three existing end-to-end autonomous driving VLMs as baselines for the BDD-X dataset: DriveGPT4 [62], RAG-Driver [65], and ADAPT [23].

Table 1: Zero-shot results on **Scenario Understanding Benchmark** provided by **Dolphins**. Only Video-LLaMA and GPT4-V take 8 frames as video input, while other methods use 16 frames as the video input. Note that all the driving-related questions on this Benchmark are unseen for all the VLMs tested, so the comparison is fair. Detailed information about our benchmark is shown in Appendix A.

Model	Is f.t. on BDD-X?	Description	Open -Voc.	Traffic Light	Weather	Scene	Time of Day	Avg
Upper performance boundary								
GPT4-V [2]	×	62.20	45.73	56.80	49.98	49.37	45.32	51.57
Human	×	57.61	80.30	87.43	-	-	-	-
Performance on unseen AV instructions								
OpenFlamingo [6]	×	5.15	31.19	44.53	23.36	20.04	16.56	23.47
Otter-Video [29]	×	13.15	22.35	38.60	24.53	24.36	15.83	23.14
Video-LLaMA [67]	×	17.17	31.40	25.49	37.09	**40.64**	28.57	30.06
Video-ChatGPT [39]	×	22.95	39.66	46.51	42.75	38.46	32.42	37.13
Dolphins-Image	×	33.23	32.00	**59.63**	34.76	39.29	**37.99**	39.48
Dolphins	✓	**40.77**	**45.96**	57.63	**50.00**	40.56	36.64	**45.26**

Metrics. For the BDD-X benchmark, we focus on the task of action and justification. We employ a suite of language metrics as our evaluation metrics, including BLEU [46], METEOR [7], and CIDEr [57]. Due to the diverse task types and complex semantic answers on the scenario understanding benchmark and DriveLM, beyond language metrics, we also introduce ChatGPT to generate evaluation scores that have been proven feasible and effective by past LVLM works [35, 39]. Furthermore, to avoid the limitations of a single metric, we follow DriveLM [1], aggregating various metrics through weighted combinations to derive the final scores. Details can be found in Appendix D.

4.2 Quantitative Results

Scene Understanding Benchmark. To have a systematic comprehension of **Dolphins**' performance, we construct a novel quantitative benchmark using the BDD100k dataset [64] to assess the model's proficiency in scenario understanding related to autonomous driving. Specifically, we manually select 400 videos from the BDD-X [28] testing split and generate 486 questions spanning six task categories: detailed description (description), open vocabulary object identification (open-voc-object), traffic light, weather, time of day, and scene. Reference answers for description, open-voc-object, and traffic light are annotated manually, while those for the remaining tasks are sourced from BDD100K annotations (These tasks can be evaluated through the accuracy metric). **Dolphins** predicts answers based on the questions and the visual input from the driving video.

We show the results in Table 1. First, with instruction-tuning on the BDD-X dataset, **Dolphins** effectively transferred the outstanding scenario understanding capability from images to the driving video domain, resulting in an overall improvement of 5.8 points. Second, despite not being trained on video instruction-following data, **Dolphins-Image** demonstrates a stronger scenario understanding

[1] https://github.com/OpenDriveLab/DriveLM/tree/main/challenge

Table 2: Zero-shot results on DriveLM demo data. All results are obtained in the same setting as Table 1.

Model	Is f.t. on DriveLM?	Description	Perception	Prediction and Planning	Avg
Upper performance boundary					
GPT4-V [2]	×	32.37	53.36	38.98	41.57
Human	×	60.81	56.95	65.90	61.22
Performance on unseen AV instructions					
OpenFlamingo [6]	×	6.86	24.92	22.39	18.06
Otter-Video [29]	×	10.60	37.76	28.33	25.56
Video-LLaMA [67]	×	11.58	42.37	32.76	28.91
Video-ChatGPT [39]	×	16.75	34.84	37.65	29.75
Dolphins-Image	×	23.52	33.45	33.10	30.02
Dolphins	×	29.08	**44.48**	41.00	38.19
Dolphins-Plus	×	**34.80**	38.61	**44.16**	**39.19**

Table 3: Results on BDD-X dataset. 'B4', 'M', and 'C' refer to BLEU-4, METEOR, CIDER. "†" represents that, due to time limitations, only 500 examples are randomly selected from the BDD-X test split for evaluation like DriveGPT4. For RAG-Driver, since we contend that using action or justification or control signal for retrieving in-context examples will cause label leakage, we only report the results of the visual search for a fair comparison.

Model			Action			Justification		
	LLM	*w ICL?*	B4↑	C↑	M↑	B4↑	C↑	M↑
w/ Fine-tune on BDD-X training split								
S2VT [58]	×	×	30.2	179.8	27.5	6.3	53.4	11.2
SAA [27]	×	×	31.8	214.8	29.1	7.1	66.1	12.2
ADAPT† [23]	×	×	**34.8**	**249.9**	**30.5**	11.2	107.2	15.0
DriveGPT4† [62]	LLaMA-2 (7B)	×	30.0	214.0	29.8	9.4	102.7	14.6
RAG-Driver [65]	Vicuna-1.5 (7B)	✓	31.2	222.1	-	7.7	83.1	-
Dolphins†	MPT (7B)	✓	30.6	223.6	28.2	**15.0**	**134.2**	**17.3**
w/o Fine-tune on BDD-X training split								
Video-LLaMA† [67]	LLaMA (7B)	×	1.7	3.9	7.5	4.7	2.8	9.3
Otter-Video† [29]	MPT (7B)	×	1.4	5.5	7.2	4.1	5.2	10.1
Dolphins-Image†	MPT (7B)	×	2.1	8.6	10.3	0.9	16.0	10.1
Dolphins-Image†	MPT (7B)	✓	**19.2**	**49.7**	**20.4**	**8.4**	**54.0**	**12.1**

ability than existing video-based VLMs, surpassing the baseline by 2.4 points. Furthermore, despite **Dolphins** having been exposed to only four types of tasks during instruction tuning, it has demonstrated a notable ability to generalize to a wide range of driving-related instructions for scene understanding. This capability is considered a significant advantage of LLMs over existing perception models in autonomous driving (AD) [14].

Evaluation on DriveLM. In evaluating **Dolphins**'s holistic understanding of perception, prediction, and planning tasks, We also evaluate **Dolphins** on DriveLM demo dataset [13]. To accurately evaluate the performance of models, we employ some strategies to extract 216 question-answer pairs from the DriveLM demo data, comprising 15 scene description examples, 135 perception examples, and 66 prediction and planning examples. Details can be found in Appendix E. Table 2 demonstrates the superior performance of our method in both description

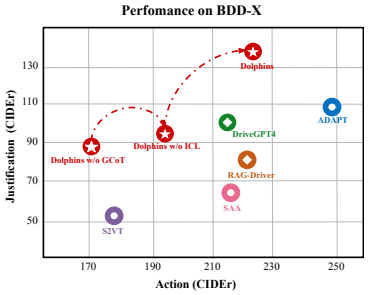

Fig. 5: Performance comparison on the test split of BDD-X. Our model achieves SOTA performance among all LVLMs [62, 65] by GCoT and adding in-context examples.

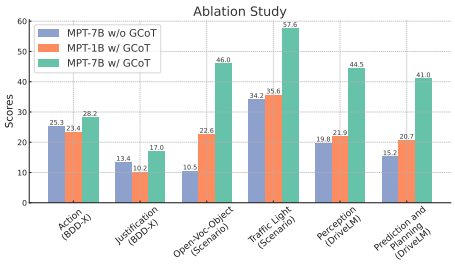

Fig. 6: Ablation study for GCoT and LLM size. "w/o GCoT" indicates that the model is not subjected to GCoT instruction tuning before finetuning on BDD-X.

and prediction and planning on DriveLM, surpassing the baseline by an overall improvement of 6.8 points. As anticipated, after instruction-tuning on the BDD-X dataset, **Dolphins** exhibits a significant improvement in overall performance on DriveLM, with an increase of 10.4 points.

Evaluation on BDD-X. In the BDD-X benchmark, we focus on the action and justification tasks due to the relevance. We follow the evaluation setting in DriveGPT4 [62] on BDD-X, but we do not provide control signals when **Dolphins** predicts vehicle action descriptions and justifications. Table 3 shows the experimental results of both zero-shot and few-shot inference on the BDD-X dataset. We also provide additional results on other tasks like control signal prediction in Appendix E.

The results demonstrate that **Dolphins** achieves superior performance compared to existing LVLMs in both action and justification prediction, showing an improvement of 1.5 points and 41.4 points on the CIDEr metric. Besides, as shown in Figure 5, introducing GCoT and in-context learning both can lead to performance improvements of 11.8% and 16.0% on the action task, respectively. However, **Dolphins** exhibits weaker performance in action description prediction compared to small expert models. Upon analyzing the results of **Dolphins** and ADAPT, we observe that models only need to learn to discern a few behaviors, such as driving forward, stopping, accelerating, and slowing down, to achieve favorable scores on BDD-X (over 200 points on the CIDEr metric). This may be attributed to an imbalance in the distribution of behaviors in the BDD-X training split. We found it is a trade-off between performance on BDDX and zero-shot performance on unseen tasks, as the training epochs increase. Detailed experiments and analysis can be found in Appendix E. Furthermore, despite **Dolphins-Image** not being fine-tuned on the BDD-X dataset, it is capable of rapid adaptation to predict actions and justifications through in-context learning. This demonstration of human-like abilities, which are anticipated to be extendable to a wide range of driving-related tasks, is particularly exciting.

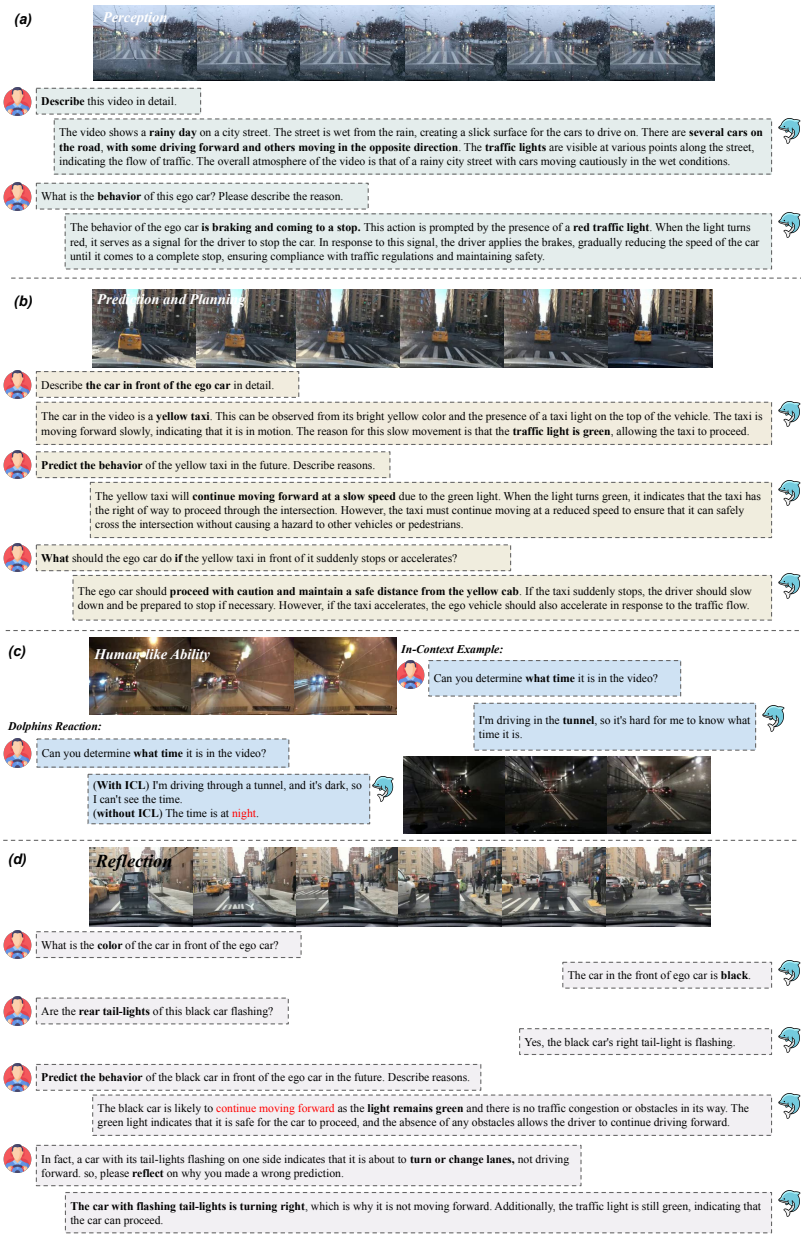

Fig. 7: The examples showing the potential of **Dolphins** to assist drivers in diverse driving scenarios through dialogue-based human-like abilities.

Ablation Study. To verify the effect of increasing the LLM size and implementing our proposed GCoT instruction tuning, we conduct an extensive ablation study,

and the results are depicted in Figure 6. We can observe that enlargement in LLM size significantly bolsters the model's generalization capabilities, culminating in substantial enhancements in zero-shot performance on tasks such as perception and planning, with an overall improvement of 22 points. Furthermore, in the absence of GCoT, the model still achieves commendable scores after being fine-tuned on the BDD-X dataset. However, its performance on unseen instructions in AD is markedly diminished.

4.3 Qualitative Demonstrations

Due to the lack of existing datasets on other desired tasks mentioned in Figure 1, we also provide qualitative demos of **Dolphins**. We showcase the capabilities of **Dolphins** across two dimensions: holistic understanding and human-like capabilities. To show the efficacy of **Dolphins**, we demonstrate its capabilities spanning over perception (Figure 7 (a)), prediction, and planning (Figure 7 (b)). For the human-like capabilities, as shown in Figure 7 (c), **Dolphins** rapidly adapts to new driving conditions. As demonstrated in Figure 7 (d), **Dolphins** effectively self-corrects based on feedback. Additional demonstrations can be found in Appendix F and on our website.

5 Conclusion and Limitation

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

Acknowledgements

We sincerely appreciate the reviewers' insightful comments and valuable feedback, which have greatly enhanced the quality of our manuscript.

References

1. Openai chat. https://chat.openai.com, accessed: 2023-10-20 6
2. Openai chat. https://openai.com/research/gpt-4v-system-card, accessed: 2023-10-20 10, 11
3. Tinychat: Large language model on the edge. https://hanlab.mit.edu/blog/tinychat, accessed: 2023-10-20 14
4. What are embeddings? https://platform.openai.com/docs/guides/embeddings/what-are-embeddings, accessed: 2023-10-20 8
5. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. ArXiv **abs/2204.14198** (2022), https://api.semanticscholar.org/CorpusID:248476411 4, 8
6. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S.Y., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P.W., Ilharco, G., Wortsman, M., Schmidt, L.: Openflamingo: An open-source framework for training large autoregressive vision-language models. ArXiv **abs/2308.01390** (2023), https://api.semanticscholar.org/CorpusID:261043320 4, 8, 9, 10, 11
7. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005) 10
8. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. arXiv preprint arXiv:1903.11027 (2019) 4
9. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023) 5
10. Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., Li, H.: End-to-end autonomous driving: Challenges and frontiers. arXiv preprint arXiv:2306.16927 (2023) 1
11. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), https://lmsys.org/blog/2023-03-30-vicuna/ 4
12. Coelho, D., Oliveira, M.: A review of end-to-end autonomous driving in urban environments. IEEE Access **10**, 75296–75311 (2022) 1
13. Contributors, D.: Drivelm: Drive on language. https://github.com/OpenDriveLab/DriveLM (2023) 2, 4, 8, 11
14. Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.D., 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) 11

15. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B.A., Fung, P., Hoi, S.C.H.: Instructblip: Towards general-purpose vision-language models with instruction tuning. ArXiv **abs/2305.06500** (2023), https://api.semanticscholar.org/CorpusID:258615266 4
16. Ding, X., Han, J., Xu, H., Zhang, W., Li, X.: Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving. arXiv preprint arXiv:2309.05186 (2023) 4, 5
17. Ding, X., Han, J., Xu, H., Liang, X., Zhang, W., Li, X.: Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. arXiv preprint arXiv:2401.00988 (2024) 4
18. Fu, D., Li, X., Wen, L., Dou, M., Cai, P., Shi, B., Qiao, Y.: Drive like a human: Rethinking autonomous driving with large language models. arXiv preprint arXiv:2307.07162 (2023) 3
19. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: 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. pp. 6904–6913 (2017) 2
20. Han, W., Guo, D., Xu, C.Z., Shen, J.: Dme-driver: Integrating human decision logic and 3d scene perception in autonomous driving. arXiv preprint arXiv:2401.03641 (2024) 4
21. Hudson, D.A., Manning, C.D.: 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. pp. 6700–6709 (2019) 2
22. Jain, A., Del Pero, L., Grimmett, H., Ondruska, P.: Autonomy 2.0: Why is self-driving always 5 years away? arXiv preprint arXiv:2107.08142 (2021) 2
23. Jin, B., Liu, X., Zheng, Y., Li, P., Zhao, H., Zhang, T., Zheng, Y., Zhou, G., Liu, J.: Adapt: Action-aware driving caption transformer. 2023 IEEE International Conference on Robotics and Automation (ICRA) pp. 7554–7561 (2023), https://api.semanticscholar.org/CorpusID:256459842 8, 9, 11
24. Jin, Y., Shen, X., Peng, H., Liu, X., Qin, J., Li, J., Xie, J., Gao, P., Zhou, G., Gong, J.: Surrealdriver: Designing generative driver agent simulation framework in urban contexts based on large language model. arXiv preprint arXiv:2309.13193 (2023) 4
25. Jordan, M.I., Mitchell, T.M.: Machine learning: Trends, perspectives, and prospects. Science **349**(6245), 255–260 (2015) 1
26. Kafle, K., Kanan, C.: An analysis of visual question answering algorithms. In: Proceedings of the IEEE international conference on computer vision. pp. 1965–1973 (2017) 2
27. Kim, J., Rohrbach, A., Darrell, T., Canny, J., Akata, Z.: Textual explanations for self-driving vehicles. In: Proceedings of the European conference on computer vision (ECCV). pp. 563–578 (2018) 11
28. Kim, J., Rohrbach, A., Darrell, T., Canny, J.F., Akata, Z.: Textual explanations for self-driving vehicles. In: European Conference on Computer Vision (2018), https://api.semanticscholar.org/CorpusID:51887402 2, 6, 8, 10
29. Li, B., Zhang, Y., Chen, L., Wang, J., Pu, F., Yang, J., Li, C., Liu, Z.: Mimic-it: Multi-modal in-context instruction tuning. ArXiv **abs/2306.05425** (2023), https://api.semanticscholar.org/CorpusID:259108295 4, 9, 10, 11
30. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. ArXiv **abs/2305.03726** (2023), https://api.semanticscholar.org/CorpusID:258547300 4, 8

31. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890 (2023) 2
32. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. ArXiv **abs/2301.12597** (2023), https://api.semanticscholar.org/CorpusID:256390509 4
33. Li, W., Pan, C., Zhang, R., Ren, J., Ma, Y., Fang, J., Yan, F., Geng, Q., Huang, X., Gong, H., et al.: Aads: Augmented autonomous driving simulation using data-driven algorithms. Science robotics **4**(28), eaaw0863 (2019) 1
34. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. ArXiv **abs/2304.08485** (2023), https://api.semanticscholar.org/CorpusID:258179774 2, 4
35. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023) 10
36. Liu, L., Lu, S., Zhong, R., Wu, B., Yao, Y., Zhang, Q., Shi, W.: Computing systems for autonomous driving: State of the art and challenges. IEEE Internet of Things Journal **8**(8), 6469–6486 (2020) 1
37. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3202–3211 (2022) 8
38. Luo, R., Zhao, Z., Yang, M., Dong, J., Qiu, M.H., Lu, P., Wang, T., Wei, Z.: Valley: Video assistant with large language model enhanced ability. ArXiv **abs/2306.07207** (2023), https://api.semanticscholar.org/CorpusID:259138706 9
39. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023) 9, 10, 11
40. Mao, J., Qian, Y., Zhao, H., Wang, Y.: Gpt-driver: Learning to drive with gpt. ArXiv **abs/2310.01415** (2023), https://api.semanticscholar.org/CorpusID:263605637 3
41. Mao, J., Ye, J., Qian, Y., Pavone, M., Wang, Y.: A language agent for autonomous driving. arXiv preprint arXiv:2311.10813 (2023) 3
42. Marcu, A.M., Chen, L., Hünermann, J., Karnsund, A., Hanotte, B., Chidananda, P., Nair, S., Badrinarayanan, V., Kendall, A., Shotton, J., et al.: Lingoqa: Video question answering for autonomous driving. arXiv preprint arXiv:2312.14115 (2023) 4
43. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 3195–3204 (2019) 2
44. Mom, G.: The evolution of automotive technology: a handbook. SAE International (2023) 1
45. Mu, Y., Zhang, Q., Hu, M., Wang, W., Ding, M., Jin, J., Wang, B., Dai, J., Qiao, Y., Luo, P.: Embodiedgpt: Vision-language pre-training via embodied chain of thought. arXiv preprint arXiv:2305.15021 (2023) 4
46. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002) 10
47. Rubin, O., Herzig, J., Berant, J.: Learning to retrieve prompts for in-context learning. arXiv preprint arXiv:2112.08633 (2021) 2, 7
48. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S.,

Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models. ArXiv **abs/2210.08402** (2022), https://api.semanticscholar.org/CorpusID:252917726 8
49. Shao, H., Hu, Y., Wang, L., Waslander, S.L., Liu, Y., Li, H.: Lmdrive: Closed-loop end-to-end driving with large language models. arXiv preprint arXiv:2312.07488 (2023) 4
50. Soori, M., Arezoo, B., Dastres, R.: Artificial intelligence, machine learning and deep learning in advanced robotics, a review. Cognitive Robotics (2023) 1
51. Tampuu, A., Matiisen, T., Semikin, M., Fishman, D., Muhammad, N.: A survey of end-to-end driving: Architectures and training methods. IEEE Transactions on Neural Networks and Learning Systems **33**(4), 1364–1384 (2020) 1
52. Team, M.N.: Introducing mpt-7b: A new standard for open-source, commercially usable llms (2023), www.mosaicml.com/blog/mpt-7b, accessed: 2023-05-05 4
53. Tian, X., Gu, J., Li, B., Liu, Y., Hu, C., Wang, Y., Zhan, K., Jia, P., Lang, X., Zhao, H.: Drivevlm: The convergence of autonomous driving and large vision-language models. arXiv preprint arXiv:2402.12289 (2024) 4
54. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models. ArXiv **abs/2302.13971** (2023), https://api.semanticscholar.org/CorpusID:257219404 4
55. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 2
56. Touvron, H., Martin, L., Stone, K.R., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D.M., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A.S., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I.M., Korenev, A.V., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models. ArXiv **abs/2307.09288** (2023), https://api.semanticscholar.org/CorpusID:259950998 4
57. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015) 10
58. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: Proceedings of the IEEE international conference on computer vision. pp. 4534–4542 (2015) 11
59. Wang, W., Xie, J., Hu, C., Zou, H., Fan, J., Tong, W., Wen, Y., Wu, S., Deng, H., Li, Z., et al.: Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. arXiv preprint arXiv:2312.09245 (2023) 4
60. Wong, K., Gu, Y., Kamijo, S.: Mapping for autonomous driving: Opportunities and challenges. IEEE Intelligent Transportation Systems Magazine **13**(1), 91–106 (2020) 2
61. Wu, D., Han, W., Wang, T., Liu, Y.H., Zhang, X., Shen, J.: Language prompt for autonomous driving. ArXiv **abs/2309.04379** (2023), https://api.semanticscholar.org/CorpusID:261660217 4

62. Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.K., Li, Z., Zhao, H.: Drivegpt4: Interpretable end-to-end autonomous driving via large language model. ArXiv **abs/2310.01412** (2023), https://api.semanticscholar.org/CorpusID:263605524 3, 4, 6, 9, 11, 12
63. Yang, S., Liu, J., Zhang, R., Pan, M., Guo, Z., Li, X., Chen, Z., Gao, P., Guo, Y., Zhang, S.: Lidar-llm: Exploring the potential of large language models for 3d lidar understanding. arXiv preprint arXiv:2312.14074 (2023) 4
64. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2636–2645 (2020) 10
65. Yuan, J., Sun, S., Omeiza, D., Zhao, B., Newman, P., Kunze, L., Gadd, M.: Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. arXiv preprint arXiv:2402.10828 (2024) 3, 4, 9, 11, 12
66. Yuan, W., Pang, R.Y., Cho, K., Sukhbaatar, S., Xu, J., Weston, J.: Self-rewarding language models. arXiv preprint arXiv:2401.10020 (2024) 9
67. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. ArXiv **abs/2306.02858** (2023), https://api.semanticscholar.org/CorpusID:259075356 9, 10, 11
68. Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Chen, K., Luo, P.: Gpt4roi: Instruction tuning large language model on region-of-interest. arXiv preprint arXiv:2307.03601 (2023) 5
69. Zhao, B., Wu, B., Huang, T.: Svit: Scaling up visual instruction tuning. arXiv preprint arXiv:2307.04087 (2023) 2
70. Zhao, H., Cai, Z., Si, S., Ma, X., An, K., Chen, L., Liu, Z., Wang, S., Han, W., Chang, B.: Mmicl: Empowering vision-language model with multi-modal in-context learning. arXiv preprint arXiv:2309.07915 (2023) 4
71. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. ArXiv **abs/2304.10592** (2023), https://api.semanticscholar.org/CorpusID:258291930 4
72. Zhu, W., Hessel, J., Awadalla, A., Gadre, S.Y., Dodge, J., Fang, A., Yu, Y., Schmidt, L., Wang, W.Y., Choi, Y.: Multimodal c4: An open, billion-scale corpus of images interleaved with text. ArXiv **abs/2304.06939** (2023), https://api.semanticscholar.org/CorpusID:258170467 8