DriveGPT4: Interpretable End-to-end Autonomous Driving via Large Language Model

Zhenhua Xu [1] **Yujia Zhang** [2] **Enze Xie** [3*] **Zhen Zhao** [4] **Yong Guo** [3]
Kwan-Yee. K. Wong [1] **Zhenguo Li** [3] **Hengshuang Zhao** [1*]
[1]The University of Hong Kong, [2]Zhejiang University, [3]Huawei Noah's Ark Lab, [4]University of Sydney

Project page: https://tonyxuqaq.github.io/projects/DriveGPT4/

Abstract

Multimodal large language models (MLLMs) have emerged as a prominent area of interest within the research community, given their proficiency in handling and reasoning with non-textual data, including images and videos. This study seeks to extend the application of MLLMs to the realm of autonomous driving by introducing DriveGPT4, a novel interpretable end-to-end autonomous driving system based on LLMs. Capable of processing multi-frame video inputs and textual queries, DriveGPT4 facilitates the interpretation of vehicle actions, offers pertinent reasoning, and effectively addresses a diverse range of questions posed by users. Furthermore, DriveGPT4 predicts low-level vehicle control signals in an end-to-end fashion. These advanced capabilities are achieved through the utilization of a bespoke visual instruction tuning dataset, specifically tailored for autonomous driving applications, in conjunction with a mix-finetuning training strategy. DriveGPT4 represents the pioneering effort to leverage LLMs for the development of an interpretable end-to-end autonomous driving solution. Evaluations conducted on the BDD-X dataset showcase the superior qualitative and quantitative performance of DriveGPT4. Additionally, the fine-tuning of domain-specific data enables DriveGPT4 to yield close or even improved results in terms of autonomous driving grounding when contrasted with GPT4-V. The code and dataset will be publicly available.

1. Introduction

Over the past decade, there has been remarkable growth in the field of autonomous driving, encompassing both academia and industry (Singh & Saini, 2021; Liu et al., 2021; Parekh et al., 2022). Commercialized autonomous driving systems have been successfully implemented in everyday scenarios, such as harbors, warehouses and urban areas. Commonly, the autonomous vehicle adopts modular designs, including perception, planning, and control. In conventional autonomous driving systems, these modules are implemented by detailed rule-based methods to handle various scenarios. But such a system may fail when unseen cases are met, such as rare accidents.

To ensure that vehicles can effectively handle diverse situations using intelligent actions, data-driven learning-based methods have become a widespread component of modern autonomous driving systems (Zhao et al., 2017; Xue et al., 2019; Xu et al., 2022; 2023a;b). To better integrate and optimize the entire system, some approaches propose training the network in an end-to-end manner, eliminating the need for discontinuous intermediate steps (Prakash et al., 2021; Hu et al., 2023; Chen et al., 2023). By using vehicle-mounted sensor data as input, the end-to-end autonomous driving system can directly predict planned paths and/or low-level vehicle controls. Nonetheless, the end-to-end learning-based autonomous driving system functions as a black box, signifying that humans cannot interpret or comprehend the generated decisions, leading to significant ethical and legal concerns.

In recent years, explainable autonomous driving (Deruyttere et al., 2019; Kim et al., 2019; Atakishiyev et al., 2021; Jin et al., 2023; Malla et al., 2023) has garnered increasing interest due to its potential to demystify the black box. These studies develop large-scale datasets comprising autonomous vehicle data along with language pairs. Language models, such as BERT (Devlin et al., 2018) and GPT (Radford et al., 2018), are trained on these datasets to generate natural language based on input from vehicle-mounted sensor data. However, the capabilities of small language models are limited, causing most of these systems to produce rigid responses to predefined questions. In addition, small language models suffer from insufficient model capacity and present unsatisfactory performance.

* Corresponding authors

With the advent of large language models (LLMs), such as ChatGPT (OpneAI, 2023) and LLaMA (Touvron et al., 2023a), the interpretability of autonomous driving systems could benefit from improved text prediction, given that LLMs possess extensive general knowledge about the world. Moreover, LLMs have the potential to better analyze and generate low-level vehicle controls due to their inherent reasoning capabilities. To achieve this, LLMs are required to comprehend multimodal data, like images or videos. Multimodal LLMs have been attracting increasing interest from various research communities, such as computer vision (Li et al., 2022b;a), embodied AI (Driess et al., 2023; Liang et al., 2023), and biomedicine (Karabacak & Margetis, 2023; Li et al., 2023a). These studies propose to project multimodal input from image, audio, video, control, and other spaces into the text domain, allowing LLMs to understand and process this multimodal data as text. To the best of our knowledge, no existing paper grounds LLMs for interpretable end-to-end autonomous driving purposes.

In this paper, we introduce DriveGPT4, an interpretable end-to-end autonomous driving system that utilizes large language models. The digit "4" in the system name represents multimodality, similar to that of MiniGPT4 (Zhu et al., 2023). DriveGPT4 takes as input a video sequence captured by a front-view monocular RGB camera, and then predicts the control signal for the next step (i.e., vehicle speed and turning angle). At the same time, human users can converse with DriveGPT4, which can provide natural language responses, such as describing the vehicle's actions and explaining the reasoning behind its behavior. To train DriveGPT4 to communicate like a human, we follow LLaVA (Liu et al., 2023) and create a visual instruction tuning dataset based on the BDD-X dataset (Kim et al., 2018) using ChatGPT. The contributions of this paper are summarized as follows:

- We present DriveGPT4, a novel multimodal LLM for interpretable end-to-end autonomous driving. Mix-finetuned on the created dataset, DriveGPT4 can process multimodal input data and generate text responses as well as low-level control signals.
- We develop a new visual instruction tuning dataset for interpretable autonomous driving with the assistance of ChatGPT. The performance of DriveGPT4 is boosted by finetuning the generated data.
- We evaluate all methods on the BDD-X dataset for multiple tasks. DriveGPT4 outperforms all baselines, which demonstrates its effectiveness.

2. Related Works

End-to-end Autonomous Driving. End-to-end autonomous driving aims to directly predict the vehicle path and low-level control signals based on visual inputs (Bojarski et al., 2016; Xiao et al., 2020; Prakash et al., 2021; Hu et al., 2023; Chen et al., 2023). (He et al., 2016) is considered the first deep learning end-to-end self-driving work. In this study, the authors train a convolutional neural network to control vehicles using monocular images as input. Recent works integrate all system modules by tokenizing module outputs (Hu et al., 2023; Chen et al., 2023), achieving a more powerful and robust control effect. However, these works lack interpretability, which limits their trustworthiness and commercialization potential.

Interpretable Autonomous Driving. To address the black box issue in learning-based autonomous driving, some studies employ visualizations (Kim & Canny, 2017; Wang et al., 2021; Saha et al., 2022). However, visual maps can be challenging for non-expert passengers to comprehend. Alternatively, other research utilizes language models to describe vehicle situations with natural languages, such as vehicle actions (Deruyttere et al., 2019; Kim et al., 2019; Jin et al., 2023), vehicle action reasoning (Jin et al., 2023), surrounding object statements (Malla et al., 2023), and discussions of potential risks to the ego vehicle (Malla et al., 2023). Constrained by the limited capacity of smaller language models, these methods can only address predefined human questions and provide inflexible answers, hindering their widespread application in real-world scenarios.

Multimodal LLM. Building on the powerful pretrained LLM weights, such as PaLM (Chowdhery et al., 2022; Driess et al., 2023), LLaMA (Touvron et al., 2023a;b), and Vicuna (Peng et al., 2023), multimodal LLMs aim to handle multiple types of input beyond text. Blip (Li et al., 2022a; 2023b) leverages Q-formers to project multimodal input into the text space, while others (Li et al., 2023a; Luo et al., 2023) simply train a fully connected layer as the projector. Multimodal LLMs have been widely applied to various tasks, such as image understanding (Li et al., 2023b; Liu et al., 2023), video understanding (Luo et al., 2023; Zhang et al., 2023; Wang et al., 2023; Zhu et al., 2023; Li et al., 2023c), medical diagnosis (Li et al., 2023a; Karabacak & Margetis, 2023), and embodied AI (Chowdhery et al., 2022; Driess et al., 2023; Brohan et al., 2023; Liang et al., 2023), etc. Our task is closely related to video understanding and embodied AI. DriveGPT4 is inspired by the former to understand input video data and the latter to predict control signals. Among these works, only a few focus on autonomous driving-related tasks (Fu et al., 2023; Wu et al., 2023; Contributors, 2023). DriveLikeHuman (Fu et al., 2023) can only handle simple simulation scenes, limiting its real-world applicability. NuPrompt (Wu et al., 2023) focuses on object tracking for vehicle perception but does not consider end-to-end driving or vehicle action reasoning. DriveLM (Contributors, 2023) is a large benchmark for driving scene understanding.

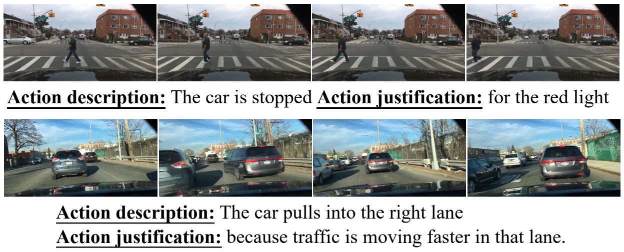


Figure 1. Example of BDD-X labeled data.

3. Data Generation

BDD-X Dataset. The BDD-X dataset (Kim et al., 2018) is employed in this study due to the scarcity of publicly available datasets suitable for our task. We sourced both videos and labels from the BDD-X dataset. This dataset contains approximately 20,000 samples, which consist of 16,803 clips designated for training and 2,123 for testing. Each clip is divided into eight images. The BDD-X dataset provides control signal data for each frame, such as vehicle speed and turning angle. It also includes text annotations detailing vehicle action descriptions and action justifications for every video clip, as exemplified in Fig. 1.

BDD-X question-answerings. BDD-X provides three types of labels: vehicle action descriptions, action justifications, and control signals for each video clip. To train the LLM, question-answering (QA) pairs are required. We generate a set of synonymous questions and use corresponding BDD-X labels as the answer. For example, for a vehicle action description, a question equivalent to *"What is the current action of this vehicle?"* should be sent to the LLM as the input question. Then, the LLM should generate the response, whose ground truth label is the vehicle action description. Considering there are three types of labels in the BDD-X dataset, we create three question sets: Q_a, Q_j, and Q_c. To prevent the LLM from overfitting to fixed question patterns, inspired by (Liu et al., 2023), each question set should contain multiple synonymous expressions of one question.

- Q_a contains synonymous questions equivalent to *"What is the current action of this vehicle?"*.
- Q_j contains synonymous questions equivalent to *"Why does this vehicle behave in this way?"*.
- Q_c contains synonymous questions equivalent to *"Predict the speed and turning angle of the vehicle in the next frame."*.

A randomly selected question $q_X \in Q_X$ and a corresponding label form a QA pair to create the dataset. LLMs can learn to predict and interpret vehicle actions simultaneously. However these QA pairs have fixed and rigid contents. Due to the lack of diversity, training solely on these QAs will degrade the ability of LLMs and render them incapable of answering questions in other formats.

Additional QAs generated by ChatGPT In previous works, ADAPT (Jin et al., 2023) trains a caption network to predict descriptions and justifications. However, the provided description and justification labels are fixed and rigid. If human users wish to learn more about the vehicle and ask everyday questions, past works may fall short. Thus, BDD-X alone is insufficient for meeting the requirements of interpretable autonomous driving. Instruction tuning data generated by ChatGPT/GPT4 has been proven effective for performance enhancement in natural language processing (Peng et al., 2023), image understanding (Liu et al., 2023), and video understanding (Li et al., 2023c; Zhang et al., 2023). ChatGPT/GPT4 can access privileged information (e.g., image-labeled captions, ground truth object bounding boxes) and is prompted to generate conversations, descriptions, and reasoning. Currently, there is no visual instruction-following dataset tailored for autonomous driving purposes. Therefore, we create our own dataset based on BDD-X assisted by ChatGPT.

To address the aforementioned issue, ChatGPT is leveraged as a teacher to generate more conversations about the ego vehicle. The prompt generally follows the prompt design used in LLaVA. To enable ChatGPT to "see" the video, YOLOv8 (Reis et al., 2023) is implemented to detect commonly seen objects in each frame of the video (e.g., vehicles, pedestrians). Obtained bounding box coordinates are normalized following LLaVA and sent to ChatGPT as privileged information. In addition to object detection results, the video clip's ground truth control signal sequences and captions are also accessible to ChatGPT. Based on this privileged information, ChatGPT is prompted to generate multiple rounds and types of conversations about the ego vehicle, traffic lights, turning directions, lane changes, surrounding objects, spatial relations between objects, etc. Detailed prompt is provided in the appendix.

Finally, we collect 56K video-text instruction-following samples, including 16K BDD-X QAs and 40K QAs generated by ChatGPT. An example of a generated sample is shown in Tab. 1.

4. DriveGPT4

4.1. Model Architecture

DriveGPT4 is a versatile multimodal LLM capable of handling various input types, including videos, and texts. Videos are uniformly sampled into a fixed number of images, and a video tokenizer based on Valley (Luo et al., 2023) is employed to convert video frames into text domain tokens. All generated tokens are concatenated and input into the LLM. In this paper, LLaMA 2 (Touvron et al., 2023b)

Caption1: Action description
The car turns right
Caption 2: Action justification
As the road is clear to turn.
Control signals
Speed(m/s): 3.91, 3.1, 2.35, 2.92, 3.51, 4.24, 4.85, 5.22
Turning angle(degree): 0.0, -3.08, -5.98, -4.66, -2.91, 1.97, 7.02, 14.46
Object boxes
Frame 0: car:[0.298, 0.408, 0.572, 0.756], car:[0.924, 0.408, 1.0, 0.51], car:[0.005, 0.83, 0.995, 0.982], car:[0.737, 0.373, 0.933, 0.522], car:[0.737, 0.373, 0.933, 0.522]
Frame 1: car:[0.327, 0.416, 0.623, 0.779], car:[0.004, 0.827, 0.99, 0.982], car:[0.961, 0.426, 1.0, 0.523], car:[0.76, 0.379, 0.966, 0.538]
Frame 2: car:[0.393, 0.427, 0.709, 0.777], car:[0.79, 0.387, 0.945, 0.553], car:[0.003, 0.825, 0.99, 0.98], car:[0.926, 0.434, 1.0, 0.549]
Frame 3: car:[0.518, 0.424, 0.849, 0.791], car:[0.834, 0.397, 0.994, 0.587], car:[0.007, 0.825, 0.983, 0.985]
Frame 4: car:[0.695, 0.542, 0.924, 0.777]
Frame 5:
Frame 6:
Frame 7:

BDD-X QAs
Question: What is the action of this vehicle?
Answer: The car turns right.
Question: Why does this vehicle behave in this way?
Answer: as the road is clear to turn.
Question: Predict the speed of the ego vehicle.
Answer: 5.5.
Question: Predict the turning angle of the ego vehicle.
Answer:7.44.

QAs generated by ChatGPT
Question: What objects are present in the video, and how do they change throughout the frames?
Answer: The video features various objects. As the video progresses, the positions and visibility of these objects change. In the initial frames, there are multiple cars. As the ego vehicle turns right, the surrounding cars gradually disappear from view.
Question: What can we learn from the ego vehicle's interactions with the traffic and surrounding environment?
Answer: The ego vehicle's interactions with the traffic and surrounding environment demonstrate the importance of safe driving practices. The driver maintains a proper speed while navigating through traffic and pays attention to their surroundings, such as the positions of other vehicles. As they make the right turn, they ensure the road is clear before accelerating. These actions highlight the need for drivers to be aware of their environment, follow traffic rules, and prioritize safety when sharing the road with other users.

Table 1. Example of the instruction-tuning data sample. The upper part of this figure demonstrates input information to ChatGPT, including video captions, control signals and object detection results obtained by YOLOv8. The lower part shows BDD-X QAs and conversations generated by ChatGPT. Refer to the appendix for detailed prompts.

is adopted as the LLM. Upon producing predicted tokens, a de-tokenizer decodes them to restore human languages. Drawing inspiration from RT-2 (Brohan et al., 2023), texts and control signals utilize the same text de-tokenizer, signifying that control signals can be interpreted as a language and effectively processed by LLMs. Decoded texts contain predicted signals in a fixed format. The overview architecture of DriveGPT4 is visualized in Fig. 2.

Video tokenizer. Let the input video frames be denoted as $V = [I_1, I_2, ..., I_N]$. For each video frame I_i, the pretrained CLIP visual encoder (Radford et al., 2021) is used to extract its feature $F_i \in \mathbb{R}^{257\times d}$. The first channel of F_i represents the global feature of I_i, while the other 256 channels correspond to patch features of I_i. For succinct representation, the global feature of I_i is denoted as F_i^G, while the local patch features of I_i are represented as F_i^P. The temporal visual feature of the entire video can then be expressed as:

$$T = F_0^G \oplus F_1^G \oplus ... \oplus F_N^G \tag{1}$$

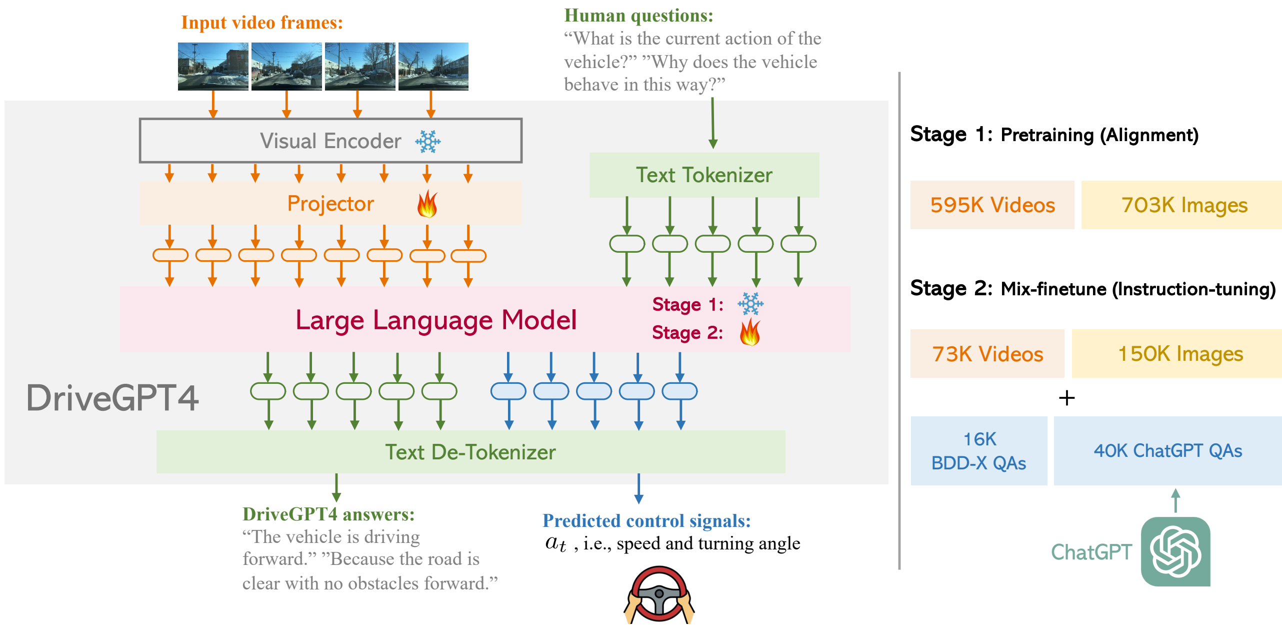


Figure 2. DriveGPT4 overview. DriveGPT4 is a comprehensive multimodal language model capable of processing inputs comprising videos, and texts. Video sequences undergo tokenization using a dedicated video tokenizer, while text and control signals share a common de-tokenizer. DriveGPT4 can concurrently generate responses to human inquiries and predict control signals.

where $\oplus$ denotes concatenation. The spatial visual feature of the whole video is given by:

$$S = \text{Pooling}(F_0^P, F_1^P, ..., F_N^P) \tag{2}$$

where Pooling$(\cdot)$ represents a pooling layer that convert N features into a single feature tensor for memory efficiency. Ultimately, both the temporal feature T and spatial feature S are projected into the text domain using a projector.

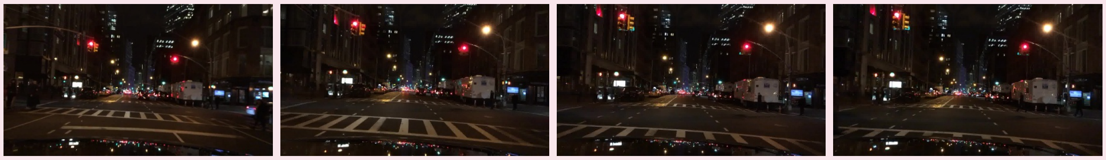
Text and control signals. Inspired by RT-2 (Brohan et al., 2023), control signals are processed similarly to texts, as they belong to the same domain space. Control signals are directly embedded within texts during the process. The default LLaMA tokenizer is employed. DriveGPT4 should predict control signals in the next step (i.e., (v_{N+1}, Δ_{N+1})) based on the multimodal input data. The time length of the input video clip and the current vehicle speed are included in the text input. The turning angle represents the relative angle between the current frame and the previous frame. After obtaining predicted tokens, the LLaMA tokenizer is used to decode tokens back into texts. Predicted control signals are embedded in the output texts using a fixed format, allowing for easy extraction. An example illustrating the input and output of DriveGPT4 is presented in Tab. 2.

4.2. Training

Consistent with previous LLM-related studies, DriveGPT4's training consists of two stages: (1) the pretraining stage, focusing on video-text alignment; and (2) the mix-finetuning stage, aimed at training the LLM to answer questions related to interpretable end-to-end autonomous driving.

Pretraining. In line with LLaVA (Liu et al., 2023) and Valley (Luo et al., 2023), the model undergoes pretraining on 593K image-text pairs from the CC3M dataset and 703K video-text pairs from the WebVid-2M dataset (Bain et al., 2021). The pretraining images and videos encompass various topics and are not specifically designed for autonomous driving applications. During this phase, the CLIP encoder and LLM weights remain fixed. Only the projector is trained.

Mix-finetune. In this stage, the LLM in DriveGPT4 is trained alongside the projector. To enable DriveGPT4 to understand and process domain knowledge, it is trained with the 56K video-text instruction-following data generated in Section 3. To enhance DriveGPT4's ability for visual understanding and question answering, 223K general instruction-following data generated by LLaVA and Valley are also utilized for mix-finetuning. "Mix" represents that general visual understanding data is utilized for training together with task-specific instruction tuning data for our task. Consequently, DriveGPT4 is finetuned with 56K video-text instruction-following data for autonomous driving in conjunction with 223K general instruction-following data. The former ensures that DriveGPT4 can be applied for interpretable end-to-end autonomous driving, while the latter enhances the data diversity and visual understanding ability of DriveGPT4.



Human: Current speed of the vehicle: 2.15m/s. The length of this video is 2.10 seconds. What is the current action of this vehicle?
Assistant: The car slows down to a stop.
Human: Why does the vehicle behave in this way?
Assistant: since the light ahead became red.
Human: Forecast the speed of the vehicle.
Assistant: 2.09.
Human: Forecast the turning angle of the vehicle.
Assistant: 0.00.

Table 2. Example of DriveGPT4 predictions. In this example, 4 out of 8 frames are shown for concise visualization.

Table 3. Testing set split.

Split	Scenes	Amount
Easy	Stopped; Driving forward; Parked; etc.	1018
Medium	Lane switch; Acceleration; Intersection; etc.	478
Hard	Vehicle turning; Traffic light change; etc.	312

5. Experiment

In this paper, DriveGPT4 focuses on interpretable end-to-end autonomous driving. With video frames and human questions as input, the method is required to predict interpretations in human language and control signals in the next step. Currently, except the BDD-X dataset, there are very few existing datasets that provide video clips captured by vehicle-mounted cameras with text interpretation and control signal annotations. Therefore, we mainly conduct evaluation experiments on the BDD-X dataset. The BDD-X dataset is filtered to remove samples that have inconsistent control signals and text reasoning.

5.1. Interpretable Autonomous Driving

In this section, we evaluate DriveGPT4 and its baselines on interpretation generation, covering vehicle action description, action justification, and additional questions about vehicle status. ADAPT (Jin et al., 2023) serves as the state-of-the-art baseline work. Recent multimodal video understanding LLMs (Zhang et al., 2023; Luo et al., 2023) are also considered for comparison. All methods use 8-frame videos as input. Currently, DriveGPT4 does not take 32-frame videos as input like ADAPT considering the heavy memory consumption and inference speed, which could be treated as a limitation of this work.

Testing Set Split. During vehicle driving, the distribution of scenes is usually not balanced. For example, some simple scenes like driving straight-forward are more commonly seen than more challenging vehicle turning or lane changes. For a comprehensive evaluation comparison, the testing set is split into "Easy", "Medium" and "Hard" sets based on the driving scene and vehicle status. Detailed split information is shown in Tab. 3.

Evaluation Metrics. To thoroughly assess the methods, we report multiple metric scores widely used in the NLP community, including CIDEr (Vedantam et al., 2015), BLEU4 (Papineni et al., 2002), and ROUGE-L (Lin, 2004). The BDD-X QA task tends to have a fixed format, so the aforementioned NLP metrics are already sufficient for evaluation. ChatGPT-generated QAs possess flexible formats and more complicated semantic meanings. Following past MLLM works (Liu et al., 2023; Li et al., 2023c; Luo et al., 2023), we also report the score generated by ChatGPT. ChatGPT is prompted to assign a numerical score between 0 and 1, with a higher score indicating better prediction accuracy. The detailed prompt for ChatGPT-based evaluation is available in the appendix. However, it should be noted that the ChatGPT score is quite not stable, thus we report the mean of three times of evaluations for qualitative reference.

Action Description and Justification. The goal is to predict vehicle action descriptions and justifications as closely as possible to the given labels. Evaluation results of all testing splits are displayed in Tab. 4. More detailed results are shown in Tab. 5. From the results, it is observed that DriveGPT4 outperforms the previous SOTA baseline ADAPT on all testing data, especially for the "Hard" splits with more challenging driving scenes and vehicle dynamics. The effectiveness and superiority of the proposed DriveGPT4 are well demonstrated.

Additional Question Answering. The above vehicle action description and justification have relatively fixed formats. To further evaluate the interpretable ability and flexibility of DriveGPT, additional questions are generated following section 3. A hundred randomly sampled video clips in the BDD-X testing set are used for question generation. Compared with action descriptions and justifications, these

Table 4. Quantitative results of comparison experiments on different splits of the BDD-X testing dataset. We provide evaluation results on comprehensive text answering (i.e., combining description and justification). "B4" represents the BLEU4 metric score.

Method	Easy			Medium			Hard			All		
	CIDEr↑	B4↑	ROUGE↑	CIDEr↑	B4↑	ROUGE↑	CIDEr↑	B4↑	ROUGE↑	CIDEr↑	B4↑	ROUGE↑
ADAPT	108.95	21.34	46.79	74.49	15.11	41.77	48.94	11.21	40.10	93.66	17.76	44.32
Video-LLaMA	10.31	2.59	11.47	9.10	1.34	9.08	2.99	1.12	9.09	8.90	1.52	10.86
Valley	31.27	5.31	44.29	17.35	4.10	31.55	9.76	2.34	20.46	20.91	4.75	14.54
DriveGPT4	**121.24**	**22.24**	**47.52**	**76.64**	**16.43**	**42.08**	**62.86**	**13.48**	**41.85**	**102.71**	**19.00**	**45.10**

Table 5. Quantitative results of comparison experiments on the whole BDD-X testing dataset. We provide evaluation results on action description, action justification, and full-text generation (i.e., i.e., combining description and justification). "B4" stands for BLEU4.

Method	Description			Justification			Full		
	CIDEr↑	B4↑	ROUGE↑	CIDEr↑	B4↑	ROUGE↑	CIDEr↑	B4↑	ROUGE↑
ADAPT	219.35	33.42	61.83	94.62	9.95	**32.01**	93.66	17.76	44.32
DriveGPT4	**254.62**	**35.99**	**63.97**	**101.55**	**10.84**	31.91	**102.71**	**19.00**	**45.10**

Table 6. Quantitative results of control signals prediction on the whole BDD-X testing dataset.

Method	Speed (m/s)					Turning angle (degree)				
	RMSE↓	$A_{0.1}\uparrow$	$A_{0.5}\uparrow$	$A_{1.0}\uparrow$	$A_{5.0}\uparrow$	RMSE↓	$A_{0.1}\uparrow$	$A_{0.5}\uparrow$	$A_{1.0}\uparrow$	$A_{5.0}\uparrow$
ADAPT	3.02	9.56	24.77	37.07	90.39	11.98	27.93	66.83	75.13	89.45
DriveGPT4	1.30	30.09	60.88	79.92	98.44	8.98	59.23	72.89	79.59	95.32

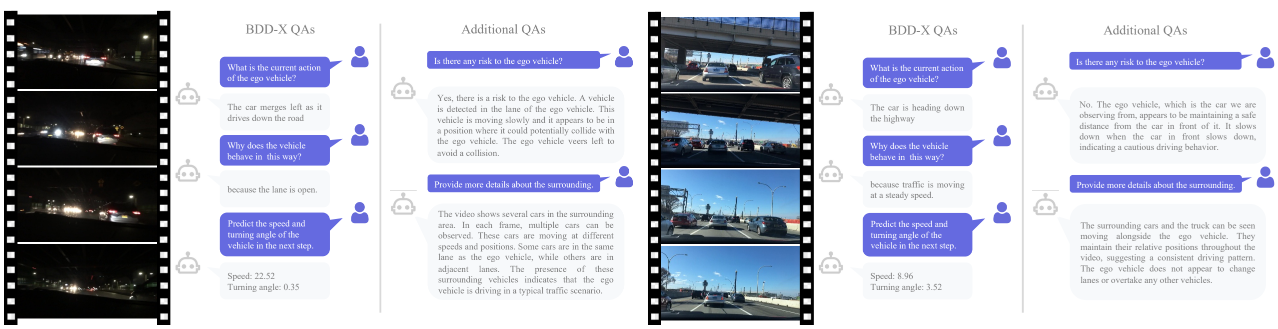


Figure 3. QAs of DriveGPT4 on the BDD-X testing set.

Table 7. Quantitative results of comparison experiments on additional question answering. The model is required to answer questions generated by ChatGPT. "B4" stands for BLEU4. "-" indicates the value is not available.

Method	CIDEr↑	B4↑	ROUGE↑	ChatGPT↑
ADAPT	-	-	-	-
Video-LLaMA	5.71	2.94	10.20	27.75
Valley	11.37	5.01	11.09	43.23
DriveGPT4	**56.34**	**22.94**	**31.70**	**81.62**

questions are more diverse and flexible. The evaluation results are shown in Tab. 7. ADAPT cannot answer additional questions except for the vehicle action description and justification. Previous video understanding LLMs can answer these questions but they do not learn autonomous driving domain knowledge. Compared with all baselines, DriveGPT4 presents superior results, demonstrating its flexibility.

5.2. End-to-end Control

In this section, we evaluate DriveGPT4 and its baselines for open-loop control signal prediction, specifically focusing on speed and turning angle. All methods are required to predict control signals for the next time step. Following previous works on control signal prediction, we use root mean squared error (RMSE) and threshold accuracies (A_τ) for evaluation. A_τ measures the proportion of test samples with prediction errors lower than τ. For a comprehensive comparison, we set τ with multiple values: $\{0.1, 0.5, 1.0, 5.0\}$. The quantitative results for the previous state-of-the-art (SOTA) method ADAPT and DriveGPT4 are shown in Tab. 6. DriveGPT4 achieves superior results for both speed and turning angle predictions.

5.3. Qualitative Results.

Multiple qualitative results are provided for intuitive comparison. For concise visualization, we only show four frames

Table 8. Quantitative results of ablation studies on the BDD-X dataset. "BQ", "CQ" and "MF" represent BDD-X QAs, ChatGPT QAs, and Mix-finetune, respectively. "-" indicates the value is not available.

BQ	CQ	MF	BDD-X Questions			ChatGPT Questions			Speed			Turning Angle		
			CIDEr↑	B4↑	ROUGE↑	CIDEr↑	B4↑	ChatGPT↑	RMSE↓	$A_{0.1}\uparrow$	$A_{1.0}\uparrow$	RMSE↓	$A_{0.1}\uparrow$	$A_{1.0}\uparrow$
			20.91	4.75	14.54	11.37	5.01	43.23	-	-	-	-	-	-
✓		✓	95.75	18.25	44.03	9.96	2.54	31.03	1.69	22.82	77.91	9.97	55.97	76.11
	✓	✓	10.40	2.31	5.35	50.09	21.53	79.37	-	-	-	-	-	-
✓	✓		76.51	11.09	40.93	17.24	10.50	77.37	4.67	17.15	44.82	14.80	21.19	70.02
✓	✓	✓	**102.71**	**19.00**	**45.10**	**56.34**	**22.94**	**81.62**	**1.30**	**30.09**	**79.92**	**8.98**	**59.23**	**79.59**

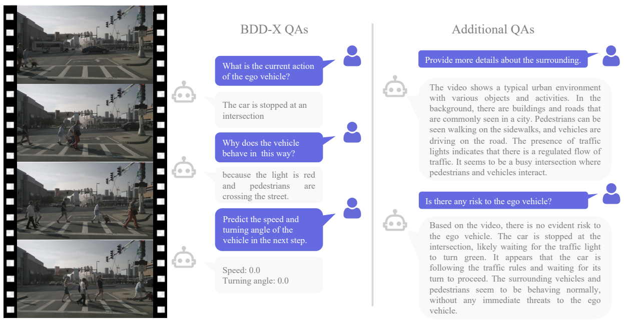


Figure 4. Zeroshot generalization of DriveGPT4 on NuScenes (Caesar et al., 2020).

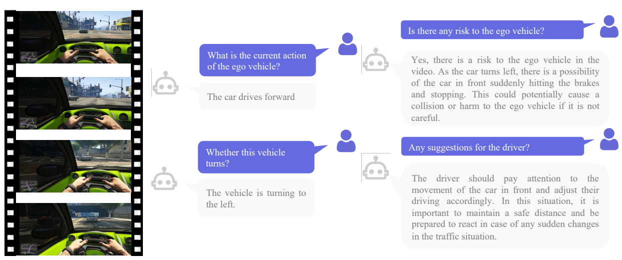


Figure 5. Zero-shot generalization of DriveGPT4 on video games.

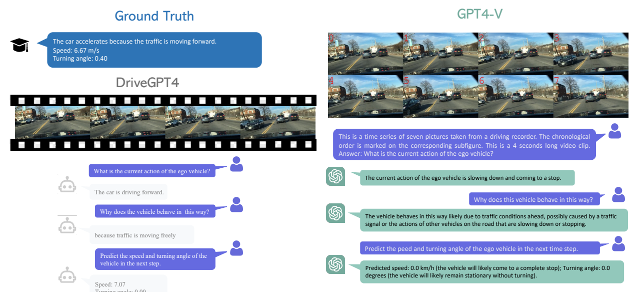


Figure 6. Comparison of DriveGPT4 and GPT4-V. GPT4-V is prompted with BDD-X QA pairs before the comparison.

of the input video clip. First, an example from the BDD-X testing set is visualized in Fig. 3. DriveGPT4 can generate high-quality texts and control predictions based on the prompt. Then, to verify the generalization ability of DriveGPT4, we apply DriveGPT4 to the NuScenes dataset (Caesar et al., 2020) for zero-shot QA in Fig. 4. We also try DriveGPT4 on video games to further test its generalization ability. An example is illustrated in Fig. 5.

GPT4-V. As the multimodal version of GPT4, GPT4-V can understand, and reason single-frame images, illustrating excellent generalization ability for various daily tasks. However, GPT4-V is still a general model for images, and not specially finetuned for grounding autonomous driving applications. Before the comparison, GPT4-V is prompted with several BDD-X QA pairs in advance. During the qualitative evaluation, even though GPT4-V illustrates powerful recognition and reasoning ability, it is observed that it (1) cannot predict numerical control signals; (2) fails to correctly understand some vehicle actions, especially dynamic actions (e.g., turning, accelerating, etc.). An example is shown in Fig. 6. More examples can be found in the appendix.

5.4. Ablation Studies

In this paper, several ablation studies are conducted to validate proposed designs, and the results are provided in Tab. 8. By removing either BDD-X QAs or ChatGPT QAs during finetuning, a decrease in corresponding performance is observed, highlighting the significance of including all task-specific multimodal data. QA pairs generated by ChatGPT enable DriveGPT4 to answer human questions in more flexible patterns, and enhance the QA ability of BDD-X questions. Then, we test DriveGPT4 without the mix-finetune strategy by removing the general image and video instruction-following data. Severe performance deduction is observed, indicating the necessity of finetuning DriveGPT4 with diverse multimodal data. Thus, changes to DriveGPT4 would negatively impact its versatile QA capabilities for interpretable end-to-end autonomous driving.

6. Conclusion

This paper presents DriveGPT4, an interpretable end-to-end autonomous driving system using multimodal LLM. A new dataset for autonomous driving interpretation is developed with the assistance of ChatGPT and employed to mix-finetune DriveGPT4, enabling it to respond to human inquiries about the vehicle. DriveGPT4 utilizes input videos and texts to generate textual responses to questions and predict control signals for vehicle operation. It outperforms baseline models in various tasks such as vehicle action description, action justification, general question answering, and control signal prediction. Moreover, DriveGPT4 exhibits generalization ability through zero-shot adaptation. In the future, DriveGPT4 will be further enhanced and adapted to close-loop vehicle control tasks.

References

Atakishiyev, S., Salameh, M., Yao, H., and Goebel, R. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *arXiv:2112.11561*, 2021.

Bain, M., Nagrani, A., Varol, G., and Zisserman, A. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., et al. End to end learning for self-driving cars. *arXiv:1604.07316*, 2016.

Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv:2307.15818*, 2023.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., and Li, H. End-to-end autonomous driving: Challenges and frontiers. *arXiv:2306.16927*, 2023.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv:2204.02311*, 2022.

Contributors, D. Drivelm: Drive on language. https://github.com/OpenDriveLab/DriveLM, 2023.

Deruyttere, T., Vandenhende, S., Grujicic, D., Van Gool, L., and Moens, M. F. Talk2car: Taking control of your self-driving car. In *EMNLP-IJCNLP*, 2019.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *ACL*, 2018.

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al. Palm-e: An embodied multimodal language model. *arXiv:2303.03378*, 2023.

Fu, D., Li, X., Wen, L., Dou, M., Cai, P., Shi, B., and Qiao, Y. Drive like a human: Rethinking autonomous driving with large language models, 2023.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.

Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al. Planning-oriented autonomous driving. In *CVPR*, 2023.

Jin, B., Liu, X., Zheng, Y., Li, P., Zhao, H., Zhang, T., Zheng, Y., Zhou, G., and Liu, J. Adapt: Action-aware driving caption transformer. *arXiv:2302.00673*, 2023.

Karabacak, M. and Margetis, K. Embracing large language models for medical applications: Opportunities and challenges. *Cureus*, 15(5), 2023.

Kim, J. and Canny, J. Interpretable learning for self-driving cars by visualizing causal attention. In *ICCV*, 2017.

Kim, J., Rohrbach, A., Darrell, T., Canny, J., and Akata, Z. Textual explanations for self-driving vehicles. *ECCV*, 2018.

Kim, J., Misu, T., Chen, Y.-T., Tawari, A., and Canny, J. Grounding human-to-vehicle advice for self-driving vehicles. In *CVPR*, 2019.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv:2306.00890*, 2023a.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022a.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ICML*, 2023b.

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

Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., et al. Grounded language-image pre-training. In *CVPR*, 2022b.

Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., and Zeng, A. Code as policies: Language model programs for embodied control. In *ICRA*, 2023.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 2004.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *NIPS*, 2023.

Liu, T., hai Liao, Q., Gan, L., Ma, F., Cheng, J., Xie, X., Wang, Z., Chen, Y., Zhu, Y., Zhang, S., et al. The role of the hercules autonomous vehicle during the covid-19 pandemic: An autonomous logistic vehicle for contactless goods transportation. *IEEE Robotics & Automation Magazine*, 2021.

Luo, R., Zhao, Z., Yang, M., Dong, J., Qiu, M., Lu, P., Wang, T., and Wei, Z. Valley: Video assistant with large language model enhanced ability. *arXiv:2306.07207*, 2023.

Malla, S., Choi, C., Dwivedi, I., Choi, J. H., and Li, J. Drama: Joint risk localization and captioning in driving. In *WACV*, 2023.

OpneAI. ChatGPT.https://openai.com/blog/chatgpt/, 2023.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

Parekh, D., Poddar, N., Rajpurkar, A., Chahal, M., Kumar, N., Joshi, G. P., and Cho, W. A review on autonomous vehicles: Progress, methods and challenges. *Electronics*, 2022.

Peng, B., Li, C., He, P., Galley, M., and Gao, J. Instruction tuning with gpt-4. *arXiv:2304.03277*, 2023.

Prakash, A., Chitta, K., and Geiger, A. Multi-modal fusion transformer for end-to-end autonomous driving. In *CVPR*, 2021.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. Improving language understanding by generative pre-training. 2018.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Reis, D., Kupec, J., Hong, J., and Daoudi, A. Real-time flying object detection with yolov8. *arXiv:2305.09972*, 2023.

Saha, A., Mendez, O., Russell, C., and Bowden, R. Translating images into maps. In *ICRA*, 2022.

Singh, S. and Saini, B. S. Autonomous cars: Recent developments, challenges, and possible solutions. In *IOP Conference Series: Materials Science and Engineering*, 2021.

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:2302.13971*, 2023a.

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

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

Wang, H., Cai, P., Sun, Y., Wang, L., and Liu, M. Learning interpretable end-to-end vision-based motion planning for autonomous driving with optical flow distillation. In *ICRA*, 2021.

Wang, J., Chen, D., Luo, C., Dai, X., Yuan, L., Wu, Z., and Jiang, Y.-G. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *arXiv:2304.14407*, 2023.

Wu, D., Han, W., Wang, T., Liu, Y., Zhang, X., and Shen, J. Language prompt for autonomous driving. *arXiv:2309.04379*, 2023.

Xiao, Y., Codevilla, F., Gurram, A., Urfalioglu, O., and López, A. M. Multimodal end-to-end autonomous driving. *TITS*, 2020.

Xu, Z., Liu, Y., Sun, Y., Liu, M., and Wang, L. Centerlinedet: Road lane centerline graph detection with vehicle-mounted sensors by transformer for high-definition map creation. *arXiv:2209.07734*, 2022.

Xu, Z., Liu, Y., Sun, Y., Liu, M., and Wang, L. Rngdet++: Road network graph detection by transformer with instance segmentation and multi-scale features enhancement. *RAL*, 2023a.

Xu, Z., Wong, K. K., and Zhao, H. Insightmapper: A closer look at inner-instance information for vectorized high-definition mapping. *arXiv:2308.08543*, 2023b.

Xue, N., Bai, S., Wang, F., Xia, G.-S., Wu, T., and Zhang, L. Learning attraction field representation for robust line segment detection. In *CVPR*, 2019.

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

Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. Pyramid scene parsing network. In *CVPR*, 2017.

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

A. Data Processing

In the data processing part, we generate three question sets for fixed question answering, i.e., Q_a for action description, Q_j for action justification and Q_c for control signals. The detailed question sets are shown in Tab. 9, Tab. 10 and Tab. 11, respectively.

The detailed prompt to generate conversations by ChatGPT is shown in Tab. 12.

What is the current action of this vehicle?
What is the vehicle doing right now in this video?
What action is the vehicle performing in this video at the moment?
Can you describe the vehicle's current activity in this video?
What's happening with the vehicle in this video right now?
At this moment in the video, what is the vehicle engaged in?
What can you observe the vehicle doing in this video currently?
How is the vehicle behaving at this point in the video?
What is the ongoing action of the vehicle in the video?
In this video, what action is the vehicle involved in at present?
Describe the current state of the vehicle in this video.

Table 9. Question set Q_a.

Why does this vehicle behave in this way?
What is the reason behind this vehicle's behavior?
Can you explain the cause of this vehicle's actions?
What factors contribute to the way this vehicle is behaving?
What's the rationale behind this vehicle's behavior?
Why is the vehicle acting in this particular manner?
What prompted the vehicle to behave like this?
What circumstances led to this vehicle's behavior?
What is the underlying cause of this vehicle's actions?
For what reason is the vehicle exhibiting this behavior?
What's driving the vehicle to behave in this way?

Table 10. Question set Q_j.

Predict the speed and turning angle of the vehicle in the next frame.
Foresee the speed and turning angle of the vehicle in the following frame.
Anticipate the speed and turning angle of the vehicle in the subsequent frame.
Estimate the speed and turning angle of the vehicle in the next frame.
Project the speed and turning angle of the vehicle in the upcoming frame.
Forecast the speed and turning angle of the vehicle in the ensuing frame.
Envision the speed and turning angle of the vehicle in the next frame.
Expect the speed and turning angle of the vehicle in the following frame.
Presume the speed and turning angle of the vehicle in the subsequent frame.
Prognosticate the speed and turning angle of the vehicle in the next frame.
Calculate the speed and turning angle of the vehicle in the upcoming frame.

Table 11. Question set Q_c.

B. Evaluation Scores Generated by ChatGPT

ChatGPT-generate QAs have long and complicated texts, thus ChatGPT score is needed. The prompt for evaluation score generation is demonstrated in Tab. 13. For each ChatGPT-generated QA pair, we embed them in the prompt texts and send them to ChatGPT. ChatGPT first outputs a numerical number ranging from 0 to 1, and then provides explanations to the score. Since ChatGPT score is not stable, we report the mean of 3 rounds of evaluations for qualitative comparison.

There is a 8-frame video recording a drive driving a vehicle. {BDD-X captions}. There are some exclusive privilege information, but you cannot mention them in your generated question answering. 1. Objects in each frame of the video: {objects}; 2. The speed (m/s) of the vehicle in each frame :{speed}. The turning angle (degree) of the vehicle in each frame :{turning angle}.

Design a conversation between you and a person asking about this video. The answers should be in a tone that a visual AI assistant is seeing the video and answering the question. Ask diverse questions and give corresponding answers.

Include questions asking about the visual content of the video, including the ego vehicle, traffic light, turning direction, lane change, surrounding objects, objects spatial relations, etc. Only include questions that have definite answers:
(1) one can see the content in the video that the question asks about and can answer confidently; (2) one can determine confidently from the video that it is not in the video. Do not ask any question that cannot be answered confidently.
Do not contain specific numbers in the questions, e.g., normalized coordinates, speed value, turning angle.

Also include complex questions that are relevant to the content in the video, for example, asking about background knowledge of the objects in the video, asking to discuss about events happening in the video, etc. Again, do not ask about uncertain details. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary.

The conversation should be 3 turns. Make the answer concise and accurate.

Table 12. Prompts for visual instruction generation. ChatGPT can access privileged information like ground truth BDD-X captions, object boxes, vehicle speeds and turning angles.

Now there are some descritions about a driver driving a vehicle. The ground truth description is: {GT label}. The description generated by deep learning model is: {Prediction}.

Give me an evaluation score about the predicted description. The score should range from 0 to 1. Larger score means better description. The score should be a float number with 2 decimal places. For example, 0.51, 0.99, 0.00, 0.76, etc.

You should first give me the score number, and then provide explanations for your score number.

Table 13. Prompts for response text evaluation. Compared with conventional metrics, the score generated by ChatGPT is more reasonable and convincing.

Differently, BDD-X QAs have a short and relatively fixed format. We did not use ChatGPT score to evaluate BDD-X QAs since ChatGPT score is quite unstable and unpredictable. For BDD-X QAs evaluation, compared with ChatGPT score, conventional metrics like CIDEr (Vedantam et al., 2015) utilized in the NLP community are more reliable.

C. Prediction Consistency

Predictions: ["The car is driving forward",
"because the light is green and there is no traffic impeding it.",
"Speed: 13.16; Turning angle: 89.05"]

ChatGPT consistent score: 0.20

Score reasoning: In this case, I would give a score of 0.2 to the consistency of the control and text. The car is driving forward because the light is green and there is no traffic impeding it, which is consistent with the given information. However, a turning angle of 89.05 indicates a sharp turn, which contradicts the statement of the car driving forward.

Predictions: ["The car is turning left",
"because the light is green.",
"Speed: 5.47; Turning angle: -2"]

ChatGPT consistent score: 0.90

Score reasoning: In this case, I would give a score of 0.9 to the consistency of the control and text. The car is turning left because the light is green, which is consistent with the given information. The speed of 5.47 and a turning angle of -2 indicate a controlled and gradual left turn.

Figure 7. Examples of consistency evaluation. In these examples, we prompt ChatGPT to explain its evaluation scores.

To measure whether DriveGPT4 makes consistent control signal and text response predictions, ChatGPT is utilized for

Here are texts describing a vehicle action and its control signals.

Positive examples:
1. [The car is stopped","because the light is red.","Speed: 0.21; Turning angle: 0.0"].
2. ["The car is turning left","because the light is green.","Speed: 5.47; Turning angle: -89.59"].

Negative examples:
1. [The car is turning left","because the light is green.","Speed: 5.47; Turning angle: 89.59"]: The car is tunring left, so that the turning angle should be negative, thus the control and the text are not consistent.
2. [The car is stopped","because the light is red.","Speed: 10.21; Turning angle: 0.0"]: The car should be stopped, but the speed is very high, which is not correct.

Now, please provide a score ranging from 0-1 to measure the consistency of the control and text of the following case. Note that you should only provide a score figure, and no other text is allowed.
{Predictions to be evaluated}

Table 14. Prompts for prediction consistency evaluation.

Table 15. Evalution results of prediction consistency.

Method	ChatGPT↑
ADAPT	79.40
DriveGPT4	80.69

consistency evaluation. ChatGPT is prompted to generate a score ranging from 0 to 1 to measure the consistency between generated text answers and predicted control signals. The detailed prompt for ChatGPT is demonstrated in Tab. 14. Consistency evaluation examples are shown in Fig. 7. The evaluation results are listed in Tab. 15. The consistency scores of ADAPT and DriveGPT4 are similar. In most situations, both ADAPT and DriveGPT4 can have consistent predictions.

D. Comparison with GPT4-V

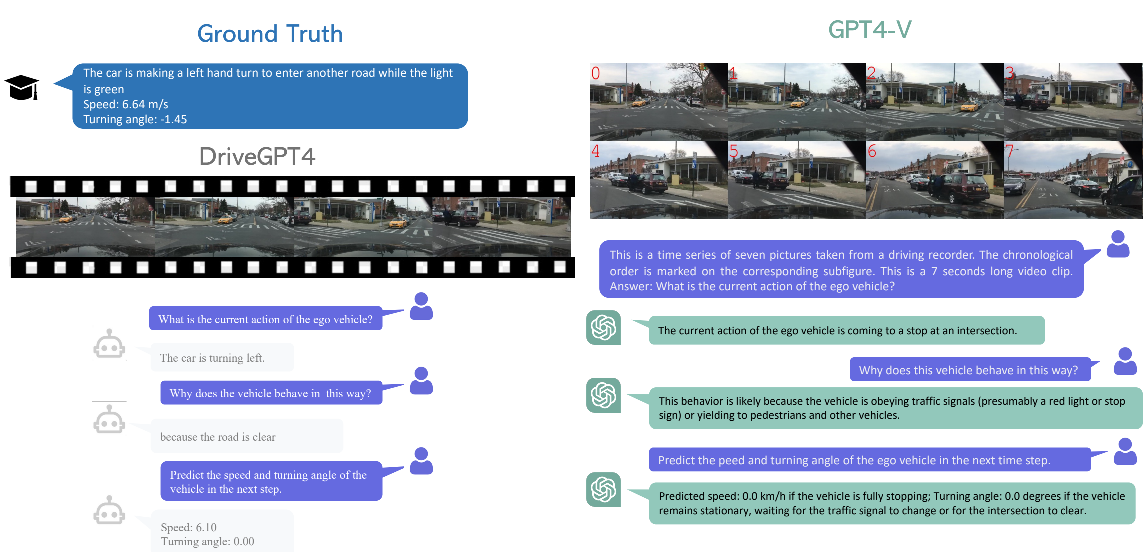


Figure 8. Comparison between GPT4-V and DriveGPT4. In this example, GPT-V fails to correctly identify the dynamics of the vehicle and cannot generate numerical control signal predictions.

As the multimodal version of GPT4, GPT4-V can understand and reason single-frame images and illustrate excellent generalization ability for various daily tasks. However, GPT4-V is still a general model for images, and not specially finetuned for grounding autonomous driving applications. Before the comparison, GPT4-V is prompted with several BDD-X

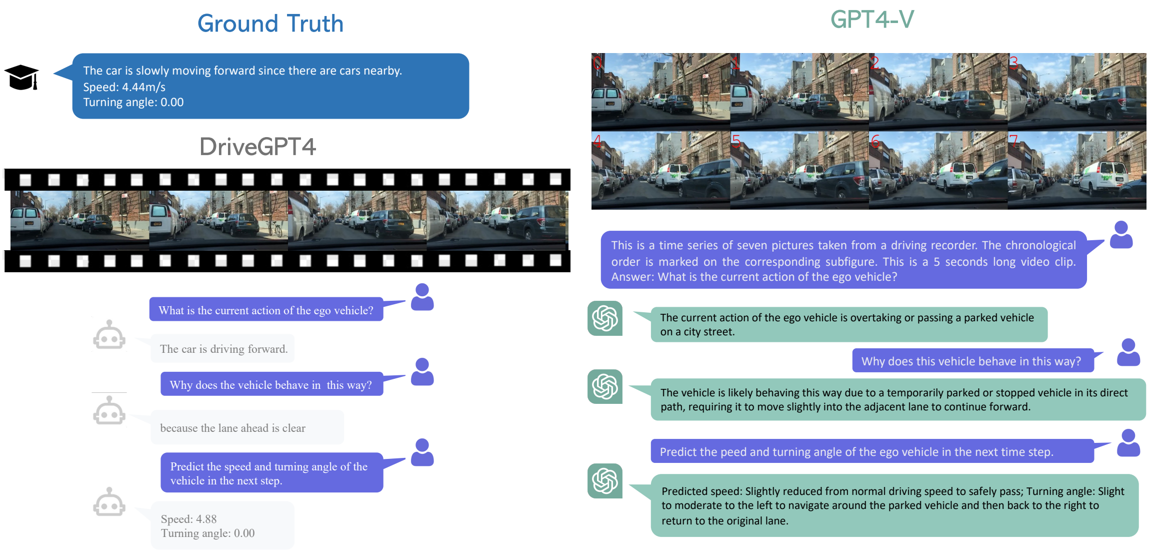


Figure 9. Comparison between GPT4-V and DriveGPT4. In this example, GPT-V fails to generate numerical control signal predictions.

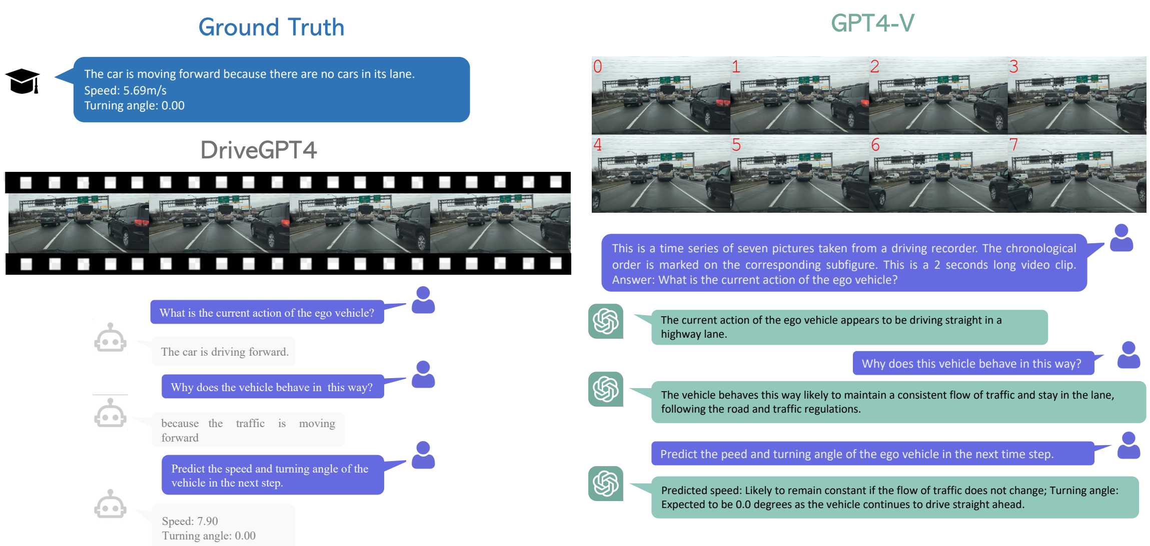


Figure 10. Comparison between GPT4-V and DriveGPT4. In this example, GPT-V cannot generate numerical control signal predictions.

QA pairs in advance. During the qualitative evaluation, even though GPT4-V illustrates powerful recognition and reasoning ability, it is observed that it (1) cannot predict numerical control signals; (2) fails to correctly understand some vehicle actions, especially dynamic actions (e.g., turning, accelerating, etc.). Examples of the comparison between GPT4-V and DriveGPT4 are shown in Fig. 8, Fig. 9 and Fig. 10.

E. Additional Qualitative Results

We provide more qualitative results of DriveGPT4 on the BDD-X dataset for interpretable end-to-end autonomous driving in Fig. 11 to 14.

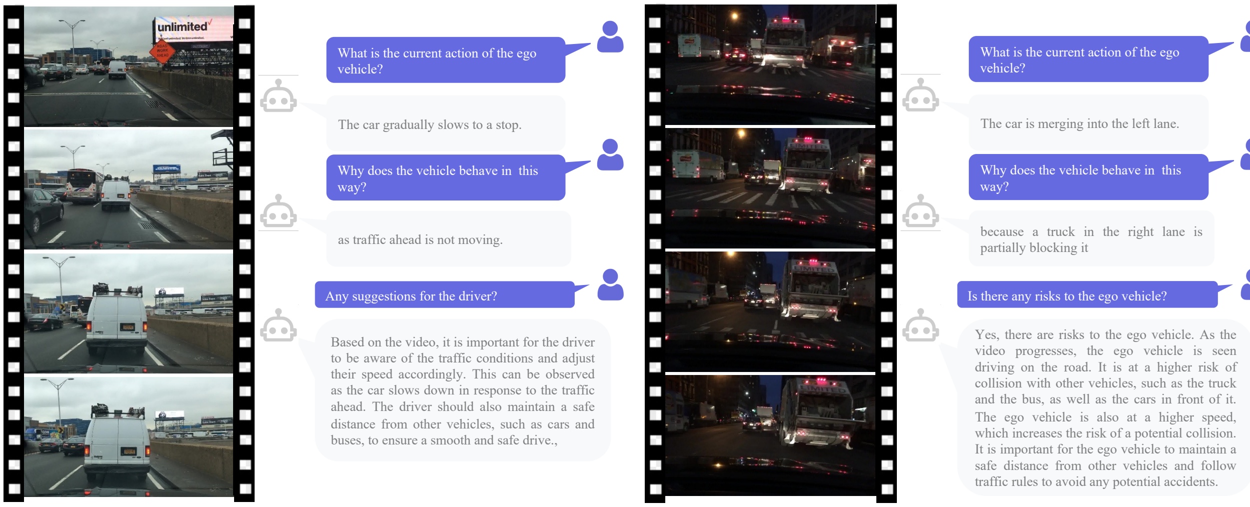


Figure 11. Conversations generated by DriveGPT4. In these examples, it is tested on the BDD-X testing set.

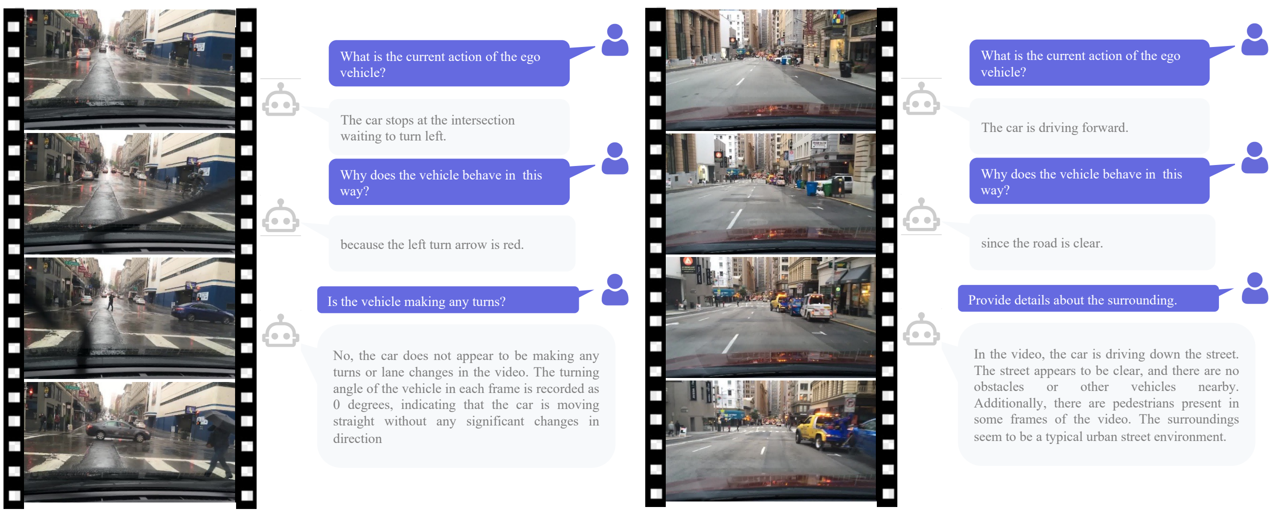


Figure 12. Conversations generated by DriveGPT4. In these examples, it is tested on the BDD-X testing set.

Figure 13. Conversations generated by DriveGPT4. In these examples, it is tested on the BDD-X testing set.

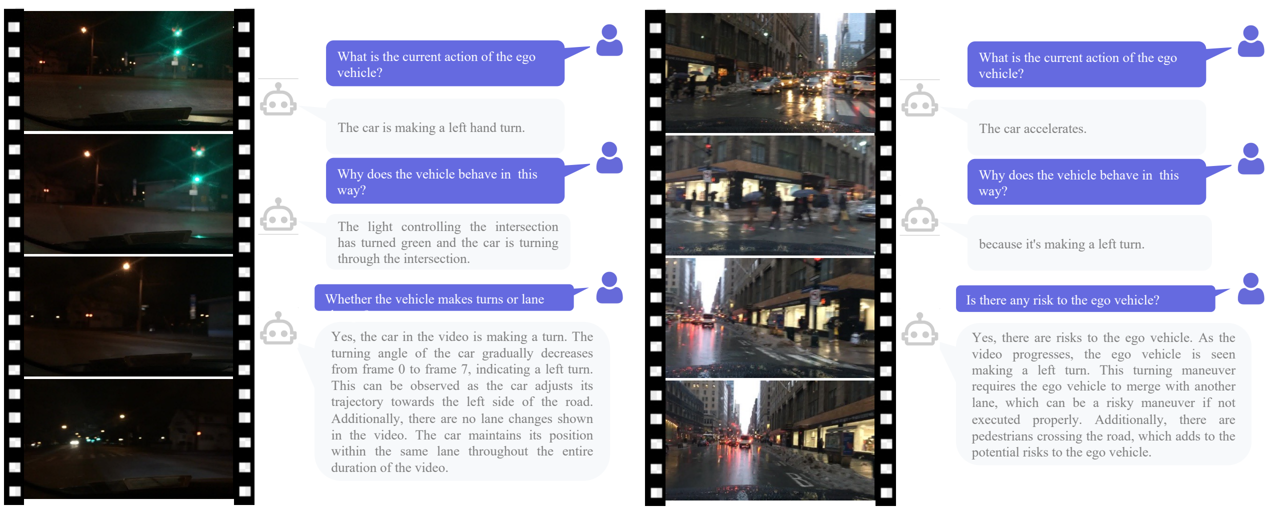


Figure 14. Conversations generated by DriveGPT4. In these examples, it is tested on the BDD-X testing set.