

AutoDrive-GPT: Enhancing Autonomous Driving Behavior Annotation and Prediction Using GPT-4o Prompt Tuning

Anonymous ICCV submission

Paper ID 8262

Abstract

001 *The rapid development of autonomous driving technology*
002 *has resulted in a substantial increase in video data gener-*
003 *ated by self-driving vehicles. Efficiently understanding and*
004 *interpreting this data is crucial for enhancing autonomous*
005 *driving systems. This paper explores the potential of GPT-*
006 *4o, a large language model, to serve as a powerful tool*
007 *for autonomous driving video tagging and reasoning. By*
008 *combining the rich video data with GPT-4o's multimodal*
009 *reasoning capabilities, we propose a structured approach,*
010 *AutoDrive-GPT, to improve autonomous driving behavior*
011 *annotation and prediction. We develop AutoDrive-GPT,*
012 *which leverages GPT-4o prompt tuning for enhanced be-*
013 *havior prediction. Additionally, we build a tool called Co-*
014 *bra that chunks video data into smaller intervals, samples*
015 *frames, and feeds them into GPT-4o for multimodal reason-*
016 *ing. Our methods are evaluated on the Bilibili and DADA-*
017 *2000 dataset, demonstrating that our approach outperforms*
018 *Gemini 1.5 flash. The results indicate that AutoDrive-GPT*
019 *significantly enhances the interpretability accuracy of au-*
020 *tonomous driving systems, particularly in challenging sce-*
021 *narios such as sudden pedestrian appearances (ghost prob-*
022 *ing) and cut-in events.*

023 1. Introduction

024 The rapid development of autonomous driving (AD) tech-
025 nology has given rise to a deluge of video data, as self-
026 driving vehicles continuously record their surroundings to
027 safely navigate complex, dynamic environments. Effi-
028 ciently interpretation of this video data remains a significant
029 challenge, as conventional video analysis methods typically
030 rely on handcrafted features or annotation-based supervised
031 learning models [3, 21, 22], which are time-consuming and
032 often fail to generalize across dynamic driving scenarios.
033 Traditional methods often focus on specific tasks, such as
034 object detection, lane line recognition, with each task typi-
035 cally handled by a separate model. This modular approach

exhibits clear difficulties when dealing with complex sce-
narios or long-tail cases, making it difficult to generalize to
unseen actions and scenarios.

Concurrently, significant progress in large language
models (LLMs) [14] and vision-language models (VLMs)
[2, 3, 5, 6, 9, 11, 17, 19], such as GPT-4o [7] and GPT-
4 [1], have demonstrated remarkable promise in address-
ing these issues. VLMs, in particular, excel at multi-
modal data interpretation, demonstrating strong capabili-
ties in action recognition, and structured output, and zero-
shot generalization [4, 14]. Their proficiency in compre-
hensively analyzing complex traffic scenes and generating
structured insights suggests that they can effectively over-
come many of the challenges associated with video caption-
ing and understanding within autonomous driving context
[12, 15, 17, 18].

By combining the rich video data generated by self-
driving vehicles with the powerful multimodal reasoning
capabilities of GPT-4o, researchers can develop robust sys-
tems for automatically tagging and annotating these video
streams. This would enable the efficient extraction of rele-
vant information, such as the identification of traffic partic-
ipants, road infrastructure, and environmental conditions,
which are essential for understanding the context and in-
forming the decision-making process of autonomous driv-
ing systems.

Leveraging these advancements, we propose an innova-
tive approach specially designed to address the limitations
of existing methods in autonomous driving video analysis.
Our methodology introduces the following key contribu-
tions:

- We propose AutoDrive-GPT, a novel automated tagging
and annotation method based on GPT-4o, capable of ef-
fectively identifying and interpreting complex and dy-
namic driving scenarios. This approach facilitates the ac-
curate and rapid extraction of critical information, such
as traffic participants movement, road infrastructure, sig-
nificantly enhancing the context-awareness and decision-
making capabilities of downstream autonomous driving
systems.

- We introduce Cobra, an efficient video processing framework that intelligently chunks and samples video data to facilitates GPT-4o analysis.
- We conduct extensive experiments using the Bibili dataset, demonstrating that our approach outperforms state-of-the-art methods across multiple metrics.
- We provide detailed analysis and insights into the capabilities and limitations of using large language models for autonomous driving applications.

Through these contributions, our work significantly advances the state-of-the-art in autonomous driving video analysis, demonstrating that the integration of sophisticated multimodal models with efficient processing frameworks can effectively meet the demands of real-world AD applications.

Our work uniquely innovates in the domain of autonomous driving video annotation by leveraging gpt-4o’s multimodal reasoning capabilities integrated with our efficient Cobra video processing framework, specifically addressing the gap in accurately recognizing rapid and safety critical driving actions, such as sudden pedestrian emergence (“ghost probing”) and abrupt lane intrusions (“cut-in”), which to our knowledge have historically posed significant difficulties for traditional video analysis methods.

2. Related Works

Interpretable Autonomous Driving. DriveGPT4 [20] is a multimodal large language model designed to integrate video-text data for enhancing both interpretability and end-to-end control in autonomous driving. DriveGPT4 utilized a fine-tuned LLaMA2 architecture combined with video-text instruction datasets to address both interpretation and control tasks in real-world driving scenarios. However, its reliance on domain-specific instruction datasets restricts its generalizability to diverse driving environments, such as surrounding vehicles or dynamic pedestrians, it only focuses on ego vehicle control.

GPT-based Motion Planner. GPT-Driver [13] is a novel approach that transforms the OpenAI GPT-3.5 model into a motion planner for autonomous driving. By reformulating motion planning as a language modelling problem, it represents planner perception input and outputs driving trajectories through language description of coordinate positions. A key innovation is the prompting-reasoning-finetuning strategy, which simulates the model’s numerical reasoning potential. The generalization and reasoning ability of GPT-3.5 enables it to tackle long-tail driving scenarios that are generally challenging to other models. In our work, we extend the GPT-based motion planner to a multimodal reasoning system that incorporates both video and audio inputs for enhanced interpretability and prediction accuracy.

Long-tail Event Detection. Long-tail event detection in autonomous driving is a challenging task due to the rarity

of certain events and the imbalanced distribution of event classes. TOKEN [16] introduces an innovative approach to handling long-tail events by tokenizing the driving environment into object-level representations. Unlike traditional end-to-end planner, TOKEN leverages a pre-trained end-to-end driving model (PARA-Drive) to generate semantically rich, object-centric tokens. Our work builds upon GPT-4o’s multimodal reasoning capabilities to enhance the interpretability and prediction accuracy of long-tail driving events, such as sudden pedestrian appearances or cut-in.

3. System Architecture

The proposed AutoDrive-GPT system consists of two main components: Cobra and GPT-4o. Cobra is responsible for processing the video data generated by autonomous vehicles, chunking the video into smaller intervals, and sampling frames evenly from each interval. These frames are then fed into GPT-4o for multimodal reasoning, where the model processes both the image and audio inputs and produces coherent text output. The system architecture is illustrated in Figure 1.

The Cobra module is primarily responsible for extracting and preprocessing multimodal information from automotive video data before this content is conveyed to the GPT-4o model for advanced reasoning. Its core functionalities are as follows: 1. Video Chunking and Frame Sampling: Cobra systematically partitions the input driving videos into smaller, temporally discrete segments. Within each chunk, it uniformly samples a predetermined number of frames. This approach preserves essential temporal and spatial information while significantly reducing computational overhead.

2. Audio Extraction and Transcription: For each temporal chunk, Cobra concurrently extracts the associated audio track and employs state-of-the-art speech-to-text services (e.g., Whisper) to generate a text transcript. This synchronized textual data augments the frame-based visual inputs, providing contextual semantic cues that enhance subsequent understanding of scene dynamics.

3. Few-shot learning and Prompt Tuning: Cobra leverages the GPT-4o model’s few-shot learning capabilities to fine-tune the multimodal reasoning process. By providing a small number of labeled examples, Cobra enables GPT-4o to rapidly adapt to new driving scenarios and predict future vehicle behaviors with high accuracy. This prompt tuning mechanism ensures that the model remains flexible and responsive to evolving driving conditions.

4. Multimodal Reasoning and Prediction: The final step in the Cobra pipeline involves feeding the processed video frames and audio transcripts into the GPT-4o model for multimodal reasoning. GPT-4o’s advanced language understanding capabilities enable it to generate coherent text outputs that summarize the observed driving behaviors and

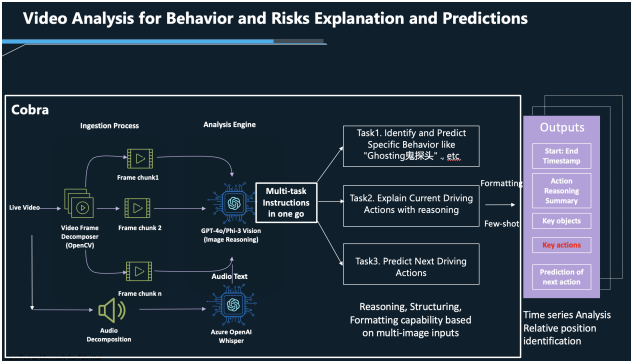


Figure 1. AutoDrive-GPT system architecture.

predict future actions. This multimodal reasoning process is crucial for enhancing the interpretability and explainability of autonomous driving systems, enabling them to make the informed driving decision.

5. Result Preservation: Cobra transmits the multimodal input bundle to GPT-4o and subsequently records the model’s JSON-formatted output. These outputs typically include action summaries detailing scene evolutions, potential hazards (e.g., sudden pedestrian appearance or abrupt lane change), and predicted vehicle actions. Storing these outputs support iterative refinement of the Autodrive-GPT system.

In essence, Cobra serves as the foundation module that bridges raw video data and sophisticated multimodal reasoning. By performing video chunking, frame sampling, audio transcription, and structured data packaging, Cobra establishes the conditions necessary for GPT-4o to deliver high-quality interpretations and predictions in complex autonomous driving scenarios. The system architecture overview is depicted in Figure 1.

4. Experiment

4.1. Dataset and Preprocessing

We evaluate the proposed AutoDrive-GPT system on the open dataset on Bilibili¹, which contains a diverse range of driving scenarios, including sudden appearances of pedestrians, lane changes, and collisions. The website consists hundreds of video clips, each having an audio commentary. We carefully selected 20 videos from the Bilibili dataset for testing. The reason we did not use a large dataset is that it is hard to find ghost probing and cut-in videos captured by front cameras of vehicles in its public videos.

To further strengthen our evaluation, we expanded the dataset by adding 80 additional videos and incorporating the DADA-2000 dataset, which provides a wider variety of challenging driving events.

¹www.bilibili.com

We compare the performance of AutoDrive-GPT with the state-of-the-art methods for autonomous driving behavior labelling and prediction. The evaluation metrics include precision, recall, and F1 score. We also conduct a qualitative analysis of the generated text outputs to assess the system’s interpretability and cohesive reasoning.

4.2. Experiment Methods

The experiment methods are as follows:

1. Data Preprocessing: We preprocess the video data using the Cobra tool, which chunks the videos into smaller intervals and samples frames evenly from each interval. We also extract audio tracks and generate text transcripts using Whisper. These frames are then fed into the GPT-4o model for multimodal reasoning.

2. Model Labelling and Reasoning: We label the video data using the GPT-4o model, extracting key driving behaviors such as ghost probing and cut-in events, then predict next actions based on the key action label, but for this experiment only key actions are evaluated in the experiment, predictions are not in the scope of evaluation.

3. Prompt Tuning [8, 10]: Prompt tuning is a crucial component of our approach, enabling GPT-4o to effectively interpret and predict driving behaviors. In this section, we detail the design of prompts, the tuning methodology. The detailed prompt tuning process is provided in Appendix A.

5. Result Analysis

The performance of the AutoDrive-GPT system was evaluated on two critical driving scenarios: cut-in and ghost probing. The results of the experiments conducted on 10 videos for each scenario are summarized in Table 1 and illustrated in Figure 3.

One example of running ghost probing labeling is shown in Figure 2.

```
report > | actionSummary-gh-ghosting-GPT.json > | s | @ scene_theme
1
2
3 {
4   "video_id": "0001",
5   "start_timestamp": "11:40",
6   "end_timestamp": "11:45",
7   "scene_theme": "ghosting",
8   "characters": "Woman in checkered dress, child in red shirt, man in black shirt",
9   "summary": "On this segment, the vehicle is driving through a narrow alley. A woman in a checkered dress is walking on the right side of the road. As the vehicle progresses, a yellow truck is seen parked ahead, partially blocking the view. At 8:40, a man in a black shirt appears from behind a white car on the right side. At 10:40, a child in a red shirt suddenly runs out from behind the white car, directly into the vehicle's path, creating a dangerous situation.",
10  "actions": "The self-driving vehicle is moving slowly through a narrow alley. The driver is advised to be cautious due to the potential for sudden pedestrian appearances. The vehicle speed is low, and it is maintaining a safe distance from the parked vehicles. The driver should be prepared to brake suddenly to avoid a collision with the child.",
11  "key_objects": "1) Right side: A woman in a checkered dress, approximately 5 meters away, walking along the road. 2) Front: A yellow truck, approximately 10 meters away, parked and partially blocking the view. 3) Right side: A white car, approximately 8 meters away, parked and blocking the view. 4) Front-right: A man in a black shirt, approximately 6 meters away, walking towards the road. 5) Front-right: A child in a red shirt, approximately 3 meters away, running into the vehicle's path.",
12  "key_actions": "ghost probing",
13  "next_action": {
14    "speed_control": "brake",
15    "direction_control": "keep direction",
16    "lane_control": "maintain current lane"
17  },
18 },
19
20 {
21   "video_id": "0002",
22   "start_timestamp": "11:40",
23   "end_timestamp": "11:45",
24   "scene_theme": "ghosting",
25   "characters": "Girl in pink jacket",
26   "summary": "On this segment, a young girl in a pink jacket is seen running across the road from the right side. The audio mentions the danger of running across the road and highlights the girl's action. The girl runs across the pedestrian crossing, narrowly avoiding a collision with a white car.",
27  "actions": "The self-driving vehicle is stationary, observing the girl running across the road. The driver comments on the danger of running across the road and advises against it. The vehicle remains stationary to avoid any potential collision.",
28  "key_objects": "1) Right side: A young girl in a pink jacket, approximately 5 meters away, running across the pedestrian crossing, directly into the vehicle's path.",
29  "key_actions": "ghost probing",
30  "next_action": {
31    "speed_control": "wait",
32    "direction_control": "keep direction",
33    "lane_control": "maintain current lane"
34  },
35 }
```

Figure 2. The result json format of running a ghost probing labelling.

Scenario	Accuracy	Recall	F1 Score	Confusion Matrix
Cut-in	0.829	0.935	0.879	{TP:29, FP:6, FN:2}
Ghost Probing	0.885	0.719	0.793	{TP:23, FP:3, FN:9}

Table 1. Final Metrics for Cut-in and Ghost Probing

Scenario	Accuracy	Recall	F1 Score	Confusion Matrix
Cut-in	0.829	0.935	0.879	{TP: 29, FP: 6, FN: 2}
Ghost Probing	0.885	0.719	0.793	{TP: 23, FP: 3, FN: 9}

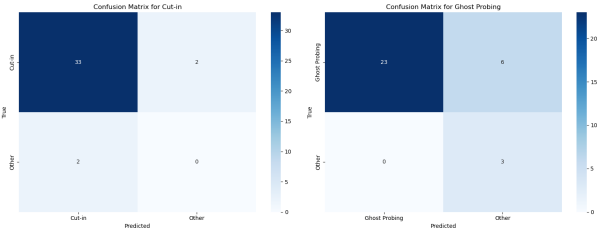


Figure 3. Confusion Matrices for Cut-in and Ghost Probing. The model performs well on the Cut-in classification with high recall and F1 score, but shows lower performance on the Ghost Probing classification with higher false positives and lower recall.

The results indicate that the AutoDrive-GPT system achieved an accuracy of 82.9% for cut-in scenarios, with a high recall of 93.5%, demonstrating its effectiveness in identifying abrupt lane changes. The F1 score of 0.879 reflects a balanced performance between precision and recall. In contrast, the ghost probing scenario yielded an accuracy of 88.5%, but the recall was lower at 71.9%, indicating challenges in detecting sudden appearances of pedestrians. The F1 score of 0.793 suggests room for improvement in this area.

The confusion matrices further elucidate the model’s performance, highlighting the true positives (TP), false positives (FP), and false negatives (FN) for each scenario. The cut-in scenario exhibited a strong performance with 29 true positives and only 2 false negatives, while the ghost probing scenario faced more challenges, with 9 false negatives indicating missed detections of pedestrians.

In summary, while the AutoDrive-GPT system demonstrates robust performance in cut-in scenarios, further refinements are necessary to enhance its detection capabilities in ghost probing situations. Future work will focus on improving the model’s sensitivity to sudden appearances of non-vehicular agents to ensure safer autonomous driving systems.

5.1. Compare gpt-4o with Gemini and Claude Sonnet 3.5

Our experimental framework included:

- **Dataset:** 80 videos from DADA-2000 (images_10_001 to images_10_080)

- **Models:** GPT-4o, Gemini-1.5-flash
- **Evaluation Metrics:** Video-level accuracy, precision, recall, and F1 score
- **Event Types:** Cut-in and ghost probing behaviors

In this section, we compare the performance of the proposed AutoDrive-GPT system with the Gemini and Claude Sonnet 3.5 models on some sample Bilibili video datasets. Gemini can analyze mp4 video format without extracting frames. The results of Gemini 1.5 flash is as follows, temperature is 0.

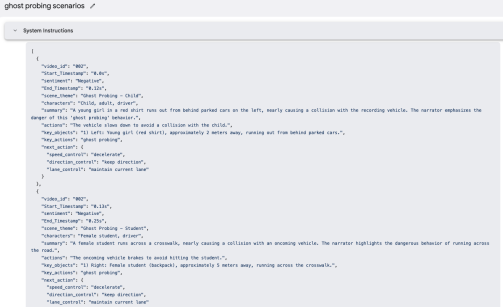


Figure 4. Gemini 1.5 flash hallucinates on start.timestamp and end.timestamp. There are only four ghost probing identified totally whereas there is only one ghost probing identified in api call so that is not posted here.

We use the same system prompt and user prompt with gpt-4o. We conduct the experiment on both Google AI Studio and api. Google AI Studio has more complete analysis than its api, whereas it still does not cover the whole ”ghost probing” during the video. Another severe problem is that Gemini 1.5 flash hallucinates on start and end timestamps and it makes hard to locate and evaluate the labels in the video. It is not recommended to continue the experiment on Gemini because the timestamp labels are incorrect and cannot be used in production.

001_ghost_probing video			
Gemini 1.5 flash		gpt-4o	
child	✓	1s-10s: child	✓
student	✓	11s-20s: girl	✓
cyclist	✓	31s-40s: cyclist	✓
child at night	✓	41s-50s: child at night	✓
		61s-70s: left-side overtaking	✓
		71s-80s: child in a t-shirt	✓

Figure 5. Gemini 1.5 flash vs GPT-4o. GPT-4o has more complete and precise analysis than Gemini 1.5 flash.

In figure 5, we can see that GPT-4o has more complete and precise analysis than Gemini 1.5 flash. The GPT-4o model can identify all the ghost probing in the video, whereas Gemini 1.5 flash can only identify four of them.

Claude 3.5 Sonnet was tested on a small set of video frames since it can only include up to 5 images for claude.ai. The api request can include up to 100 images but is unavail-

able for the author’s region, so the results of Claude 3.5 Sonnet is not posted here.

6. Conclusion

This study investigated the use of GPT-4o for enhancing video analysis in autonomous driving. We introduced AutoDrive-GPT, which utilizes GPT-4o for behavior prediction, and developed Cobra to preprocess video data for multimodal reasoning. Our evaluation on the Bilibili dataset shows that AutoDrive-GPT surpasses Gemini 1.5 Flash in terms of clarity and completeness, especially in detecting sudden pedestrian appearances and cut-in events.

Future research will aim to improve the model’s responsiveness to dynamic environments and broaden the dataset to cover more varied driving scenarios. Additionally, enhancing the mathematical reasoning capabilities of the motion planner by using gpt-o1/o3 model for trajectory inference is planned.

In summary, AutoDrive-GPT marks a significant step forward in applying large language models to autonomous driving, enhancing both safety and operational efficiency.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangoei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, pages 23716–23736. Curran Associates, Inc., 2022. 1
- [3] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19358–19369, Vancouver, BC, Canada, 2023. IEEE. 1
- [4] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai, Botian Shi, and Yu Qiao. Drive Like a Human: Rethinking Autonomous Driving with Large Language Models, 2023. *arXiv:2307.07162 [cs]*. 1
- [5] Haoxiang Gao and Yu Zhao. Application of Vision-Language Model to Pedestrians Behavior and Scene Understanding in Autonomous Driving, 2025. *arXiv:2501.06680 [cs]*. 1
- [6] Wenyi Hong, Weihang Wang, Qingsong Lv, Jiazheng Xu, Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao Dong, Ming Ding, and Jie Tang. CogAgent: A Visual Language Model for GUI Agents. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14281–14290, Seattle, WA, USA, 2024. IEEE. 1
- [7] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1
- [8] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual Prompt Tuning, 2022. *arXiv:2203.12119 [cs]*. 3
- [9] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. pages 3128–3137, 2015. 1
- [10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3
- [11] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. DeepSeek-VL: Towards Real-World Vision-Language Understanding, 2024. *arXiv:2403.05525 [cs]*. 1
- [12] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and Chaowei Xiao. Dolphins: Multimodal Language Model for Driving. In *Computer Vision – ECCV 2024*, pages 403–420, Cham, 2025. Springer Nature Switzerland. 1
- [13] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint arXiv:2310.01415*, 2023. 2
- [14] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction Tuning with GPT-4, 2023. *arXiv:2304.03277 [cs]*. 1
- [15] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. DriveLM: Driving with Graph Visual Question Answering, 2025. *arXiv:2312.14150 [cs]*. 1
- [16] Ran Tian, Boyi Li, Xinshuo Weng, Yuxiao Chen, Edward Schmerling, Yue Wang, Boris Ivanovic, and Marco Pavone. Tokenize the world into object-level knowledge to address long-tail events in autonomous driving. *arXiv preprint arXiv:2407.00959*, 2024. 2
- [17] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. DriveVLM: The Convergence of Autonomous Driving and Large Vision-Language Models, 2024. *arXiv:2402.12289 [cs]*. 1
- [18] Joseph Raj Vishal, Divesh Basina, Aarya Choudhary, and Bharatesh Chakravarthi. Eyes on the Road: State-of-the-Art Video Question Answering Models Assessment for Traffic Monitoring Tasks, 2024. *arXiv:2412.01132 [cs]*. 1
- [19] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong,

415 Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and
416 Chong Ruan. DeepSeek-VL2: Mixture-of-Experts Vision-
417 Language Models for Advanced Multimodal Understanding,
418 2024. arXiv:2412.10302 [cs]. 1

419 [20] Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo,
420 Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao.
421 Drivegpt4: Interpretable end-to-end autonomous driving via
422 large language model. *IEEE Robotics and Automation Let-*
423 *ters*, 2024. 2

424 [21] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying
425 Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Dar-
426 rell. BDD100K: A Diverse Driving Dataset for Heteroge-
427 neous Multitask Learning. pages 2636–2645, 2020. 1

428 [22] Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan
429 Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang,
430 Junkang Wu, Xue Wang, Yibo Hu, Bin Wen, Fan Yang,
431 Zhang Zhang, Tingting Gao, Di Zhang, Liang Wang, Rong
432 Jin, and Tieniu Tan. MM-RLHF: The Next Step Forward in
433 Multimodal LLM Alignment, 2025. arXiv:2502.10391 [cs].
434 1

435 **A. Prompt Tuning Details**

436 There are several prompt tuning strategies that can be used
437 to enhance the performance of GPT-4o in the context of au-
438 tonomous driving behavior prediction. These strategies in-
439 clude:

440 **A.1. multi-image video input**

441 GPT-4o supports up to 20 images input, enabling it to
442 process a sequence of images extracted from video data.
443 This capability allows the model to analyze temporal
444 changes and extract meaningful information from consec-
445 utive frames.

446 **A.2. Chain-of-Thought (CoT) Reasoning**

447 Chain-of-Thought (CoT) reasoning is a technique that en-
448 ables GPT-4o to perform complex reasoning tasks by break-
449 ing down the problem into a series of intermediate steps.
450 This approach allows the model to handle multi-step rea-
451 soning processes more effectively, improving its ability to
452 interpret and predict driving behaviors in complex scenar-
453 ios. By explicitly modeling the sequence of reasoning steps,
454 CoT enhances the model’s interpretability and accuracy in
455 decision-making.

456 1) Thought1: You are VideoAnalyzerGPT analyzing a
457 series of SEQUENTIAL images taken from a video

458 2) Thought2: Focus on the changes in the relative posi-
459 tions, distances, and speeds of objects, particularly the car
460 in front

461 3) Thought3: Pay special attention to any signs of decel-
462 eration or closing distance between the car in front and the
463 observer vehicle.

464 4) Thought4: Describe any changes in the car’s speed,
465 distance from the observer vehicle, and how these might
466 indicate a potential need for braking or collision avoidance.
467 5) Thought5: Based on the sequence of images, predict
468 the next action that the observer vehicle should take.
469 6) Thought6: If the car ahead is decelerating and the dis-
470 tance is closing rapidly, suggest whether braking is neces-
471 sary to avoid a collision.
472 7) Thought7: Examine the sequential images for visual
473 cues...Consider how these cues change from one frame to
474 the next, and describe the need for the observer vehicle to
475 take action, such as braking, based on these changes.

476 **A.3. Image-Based Few-Shot Learning**

477 Image-based few-shot learning enhances the ability to learn
478 spatial relative positions, augmenting in cross-modal align-
479 ment. For example:
480 Below are time series example images and their corre-
481 sponding analysis to help you understand how to analyze
482 and label the images:

{fsl_base64_payload} -> {assistant_response}

485 **A.4. Structured Output**

486 GPT-4o supports structured output, enabling it to gener-
487 ate well-organized and formatted responses. This capa-
488 bility is particularly useful for generating JSON, XML,
489 or other structured data formats, which can be easily
490 parsed and utilized by downstream applications. We use
491 json format and specify key columns like "key_actions",
492 "start_timestamp", "end_timestamp", and "next_actions" in
493 the output.

494 **A.5. Multi-task prompting**

495 ****Task 1: Identify and Predict potential very near future**
496 **time "Ghosting" Cut-in,etc Behavior****
497 ****Task 2: Explain Current Driving Actions****
498 ****Task 3: Predict Next Driving Action****

499 **A.6. Position-guided text prompting**

500 In order to guide the model to understand relative position
501 of objects in the image, we tell the gpt-4o model to know its
502 observing position:
503 Assume the viewpoint is standing from at the bottom
504 center of the image. Describe whether the objects are on
505 the left or right side of this central point.
506 The full prompt is in the attached code.