

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

Anonymous ICCV submission

Paper ID 8262

Abstract

The rapid development of autonomous driving technology has resulted in a substantial increase in video data generated by self-driving vehicles. Efficiently understanding and interpreting this data is crucial for enhancing autonomous driving systems. This paper explores the potential of GPT-4o, a large language model, to serve as a powerful tool for autonomous driving video tagging and reasoning. By combining the rich video data with GPT-4o's multimodal reasoning capabilities, we propose a structured approach, AutoDrive-GPT, to improve autonomous driving behavior annotation and prediction. We develop AutoDrive-GPT, which leverages GPT-4o prompt tuning for enhanced behavior prediction. Additionally, we build a tool called Cobra that chunks video data into smaller intervals, samples frames, and feeds them into GPT-4o for multimodal reasoning. Our methods are evaluated on the Bilibili and DADA-2000 dataset, demonstrating that our approach outperforms Gemini 1.5 flash. The results indicate that AutoDrive-GPT significantly enhances the interpretability accuracy of autonomous driving systems, particularly in challenging scenarios such as sudden pedestrian appearances (ghost probing) and cut-in events.

1. Introduction

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

exhibits clear difficulties when dealing with complex scenarios 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 addressing these issues. VLMs, in particular, excel at multimodal data interpretation, demonstrating strong capabilities in action recognition, and structured output, and zero-short generalization [4, 14]. Their proficiency in comprehensively analyzing complex traffic scenes and generating structured insights suggests that they can effectively overcome many of the challenges associated with video captioning 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 systems for automatically tagging and annotating these video streams. This would enable the efficient extraction of relevant information, such as the identification of traffic participants, road infrastructure, and environmental conditions, which are essential for understanding the context and informing the decision-making process of autonomous driving systems.

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

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

- 076 • We introduce Cobra, an efficient video processing frame-
077 work that intelligently chunks and samples video data to
078 facilitate GPT-4o analysis.
079 • We conduct extensive experiments using the Bibili
080 dataset, demonstrating that our approach outperforms
081 state-of-the-art methods across multiple metrics.
082 • We provide detailed analysis and insights into the capa-
083 bilities and limitations of using large language models for
084 autonomous driving applications.

085 Through these contributions, our work significantly ad-
086 vances the state-of-the-art in autonomous driving video
087 analysis, demonstrating that the integration of sophisticated
088 multimodal models with efficient processing frameworks
089 can effectively meet the demands of real-world AD appli-
090 cations.

091 Our work uniquely innovates in the domain of au-
092 tonomous driving video annotation by leveraging gpt-4o’s
093 multimodal reasoning capabilities integrated with our effi-
094 cient Cobra video processing framework, specifically ad-
095 dressing the gap in accurately recognizing rapid and safety
096 critical driving actions, such as sudden pedestrian emer-
097 gence (“ghost probing”) and abrupt lane intrusions (“cut-
098 in”), which to our knowledge have historically posed sig-
099 nificant difficulties for traditional video analysis methods.

100 2. Related Works

101 **Interpretable Autonomous Driving.** DriveGPT4 [20] is
102 a multimodal large language model designed to integrate
103 video-text data for enhancing both interpretability and end-
104 to-end control in autonomous driving. DriveGPT4 utilized a
105 fine-tuned LLaMA2 architecture combined with video-text
106 instruction datasets to address both interpretation and con-
107 trol tasks in real-world driving scenarios. However, its re-
108 liance on domain-specific instruction datasets restricts its
109 generalizability to diverse driving environments, such as
110 surrounding vehicles or dynamic pedestrians, it only fo-
111 cuses on ego vehicle control.

112 **GPT-based Motion Planner.** GPT-Driver [13] is a novel
113 approach that transforms the OpenAI GPT-3.5 model into
114 a motion planner for autonomous driving. By reformu-
115 lating motion planning as a language modelling problem,
116 it represents planner perception input and outputs driv-
117 ing trajectories through language description of coordinate
118 postions. A key innovation is the prompting-reasoning-
119 finetuning strategy, which simulates the model’s numerical
120 reasoning potential. The generalization and reasoning abil-
121 ity of GPT-3.5 enables it to tackle long-tail driving scenarios
122 that are generally challenging to other models. In our work,
123 we extend the GPT-based motion planner to a multimodal
124 reasoning system that incorporates both video and audio in-
125 puts for enhanced interpretability and prediction accuracy.

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

of certain events and the imbalanced distribution of event
128 classes. TOKEN [16] introduces an innovative approach
129 to handling long-tail events by tokenizing the driving en-
130 vironment into object-level representations. Unlike tradi-
131 tional end-to-end planner, TOKEN leverages a pre-trained
132 end-to-end driving model (PARA-Drive) to generate seman-
133 tically rich, object-centric tokens. Our work builds upon
134 GPT-4o’s multimodal reasoning capabilities to enhance the
135 interpretability and prediction accuracy of long-tail driving
136 events, such as sudden pedestrian appearances or cut-in.

138 3. System Architecture

139 The proposed AutoDrive-GPT system consists of two main
140 components: Cobra and GPT-4o. Cobra is responsible for
141 processing the video data generated by autonomous vehi-
142 cles, chunking the video into smaller intervals, and sam-
143 pling frames evenly from each interval. These frames are
144 then fed into GPT-4o for multimodal reasoning, where the
145 model processes both the image and audio inputs and pro-
146 duces coherent text output. The system architecture is illus-
147 trated in Figure 1.

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

159 2. Audio Extraction and Transcription: For each tempo-
160 ral chunk, Cobra concurrently extracts the associated au-
161 dio track and employs state-of-the-art speech-to-text ser-
162 vices (e.g., Whisper) to generate a text transcript. This syn-
163 chronized textual data augments the frame-based visual in-
164 puts, providing contextual semantic cues that enhance sub-
165 sequent understanding of scene dynamics.

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

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

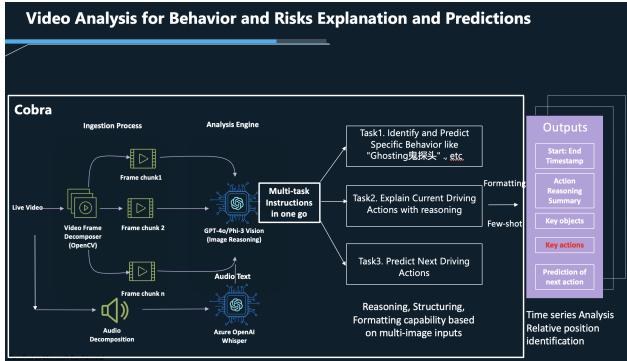


Figure 1. AutoDrive-GPT system architecture.

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.

```
(1) actionSummary-bill-ghosting-001.json <-
report:3 (1) actionSummary-bill-ghosting-001.json > (1) s > (2) scene_theme
1 {
2   (
3     "video_id": "0001",
4     "start_timestamp": "1.0s",
5     "end_timestamp": "10.0s",
6     "frame": "Frame chunk 1",
7     "characters": "Woman in checked dress, child in red shirt, man in black shirt",
8     "summary": "In this segment, the vehicle is driving through a narrow alley. A woman in a checked dress is walking on the right side of the road. As the vehicle approaches, a child in a red shirt suddenly runs from behind the white car, directly into the vehicles path, creating a dangerous situation.",
9     "actions": "The vehicle speeds up to avoid hitting the child, creating a safe distance from the passing vehicles. The driver comments on the need to brake suddenly to avoid a collision with the child.",
10    "key_actions": "ghost probing",
11    "next_actions": {
12      "action": "stop",
13      "direction": "keep direction",
14      "lane_control": "maintain current lane"
15    }
16  },
17  {
18    "video_id": "0001",
19    "start_timestamp": "11.0s",
20    "end_timestamp": "20.0s",
21    "frame": "Frame chunk 2",
22    "characters": "Man in a black shirt",
23    "summary": "A man in a black shirt appears from behind a white car on the right side. At 18s, a child in a red shirt suddenly runs from behind the white car, directly into the vehicles path, creating a dangerous situation.",
24    "actions": "The vehicle speeds up to avoid hitting the child, creating a safe distance from the passing vehicles. The driver comments on the need to brake suddenly to avoid a collision with the child.",
25    "key_actions": "ghost probing",
26    "next_actions": {
27      "action": "stop",
28      "direction": "keep direction",
29      "lane_control": "maintain current lane"
30    }
31  },
32  {
33    "video_id": "0001",
34    "start_timestamp": "21.0s",
35    "end_timestamp": "30.0s",
36    "frame": "Frame chunk n",
37    "characters": "Girl in pink jacket",
38    "summary": "In 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 advises against it. The vehicle remains stationary, narrowly avoiding a collision with a white car.",
39    "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.",
40    "key_actions": "ghost probing",
41    "next_actions": {
42      "action": "stop",
43      "direction": "keep direction",
44      "lane_control": "maintain current lane"
45    }
46  },
47 }
```

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

Video Analysis for Behavior and Risks Explanation and Predictions	215
Cobra	216
Ingestion Process	217
Analysis Engine	218
GPT-4o(Phi-3 Vision (Image Reasoning))	219
Task1. Identify and Predict Specific Behavior like "Ghosting鬼探头" , etc	220
Task2. Explain Current Driving Actions with reasoning	221
Task3. Predict Next Driving Actions	222
Formatting	223
Few-shot	224
Reasoning, Structuring, Formatting capability based on multi-image inputs	225
Outputs	226
Start-End Timestamping	227
Action Reasoning Summary	228
Key objects	229
Key actions	230
Prediction of next action	231
Time series Analysis	232
Relative position identification	233

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}

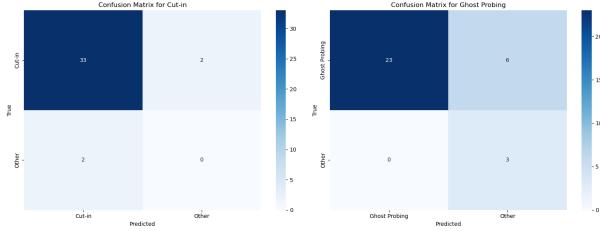


Figure 3. Confusion Matrices for Cut-in and Ghost Probing. The model performs well on the Cut-in classification with high recall and F1 score, but shows lower performance on the Ghost Probing classification with higher false positives and lower recall.

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.

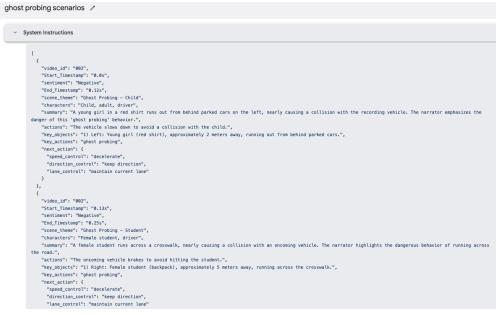


Figure 4. Gemini 1.5 flash hallucinates on start_timestamp and end_timestamp. There are only four ghost probing identified totally whereas there is only one ghost probing identified in api call so that is not posted here.

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	✓
student	✓
cyclist	✓
child at night	✓
1s-10s: child	✓
11s-20s: girl	✓
31s-40s: cyclist	✓
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 unavailable.

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

307 6. Conclusion

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

316 Future research will aim to improve the model's respon-
317 siveness to dynamic environments and broaden the dataset
318 to cover more varied driving scenarios. Additionally, en-
319 hancing the mathematical reasoning capabilities of the mo-
320 tion planner by using gpt-o1/o3 model for trajectory infer-
321 ence is planned.

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

325 References

- 326 [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ah-
327 mad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida,
328 Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al.
329 Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*,
330 2023. 1
- 331 [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine
332 Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Men-
333 sch, Katherine Millican, Malcolm Reynolds, Roman Ring,
334 Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong,
335 Sina Samangoeei, Marianne Monteiro, Jacob L Menick,
336 Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sa-
337 hand Sharifzadeh, Mikolaj Bińkowski, Ricardo Barreira,
338 Oriol Vinyals, Andrew Zisserman, and Karén Simonyan.
339 Flamingo: a Visual Language Model for Few-Shot Learn-
340 ing. In *Advances in Neural Information Processing Systems*,
341 pages 23716–23736. Curran Associates, Inc., 2022. 1
- 342 [3] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu,
343 Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue
344 Cao. EVA: Exploring the Limits of Masked Visual Re-
345 presentation Learning at Scale. In *2023 IEEE/CVF Conference
346 on Computer Vision and Pattern Recognition (CVPR)*, pages
347 19358–19369, Vancouver, BC, Canada, 2023. IEEE. 1
- 348 [4] Daocheng Fu, Xin Li, Licheng Wen, Min Dou, Pinlong Cai,
349 Botian Shi, and Yu Qiao. Drive Like a Human: Rethinking
350 Autonomous Driving with Large Language Models, 2023.
351 *arXiv:2307.07162* [cs]. 1
- 352 [5] Haoxiang Gao and Yu Zhao. Application of Vision-
353 Language Model to Pedestrians Behavior and Scene Under-
354 standing in Autonomous Driving, 2025. *arXiv:2501.06680*
355 [cs]. 1
- 356 [6] Wenyi Hong, Weihan Wang, Qingsong Lv, Jiazheng Xu,
357 Wenmeng Yu, Junhui Ji, Yan Wang, Zihan Wang, Yuxiao

Dong, Ming Ding, and Jie Tang. CogAgent: A Visual Lan-
358 guage Model for GUI Agents. In *2024 IEEE/CVF Confer-
359 ence on Computer Vision and Pattern Recognition (CVPR)*,
360 pages 14281–14290, Seattle, WA, USA, 2024. IEEE. 1

- 361 [7] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perel-
362 man, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda,
363 Alan Hayes, Alec Radford, et al. Gpt-4o system card.
364 *arXiv preprint arXiv:2410.21276*, 2024. 1
- 365 [8] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie,
366 Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Vi-
367 sual Prompt Tuning, 2022. *arXiv:2203.12119* [cs]. 3
- 368 [9] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic align-
369 ments for generating image descriptions. pages 3128–3137,
370 2015. 1
- 371 [10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee.
372 Visual instruction tuning. *Advances in neural information
373 processing systems*, 36:34892–34916, 2023. 3
- 374 [11] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong,
375 Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao
376 Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda
377 Xie, and Chong Ruan. DeepSeek-VL: Towards Real-World
378 Vision-Language Understanding, 2024. *arXiv:2403.05525*
379 [cs]. 1
- 380 [12] Yingzi Ma, Yulong Cao, Jiachen Sun, Marco Pavone, and
381 Chaowei Xiao. Dolphins: Multimodal Language Model
382 for Driving. In *Computer Vision – ECCV 2024*, pages 403–
383 420, Cham, 2025. Springer Nature Switzerland. 1
- 384 [13] Jiageng Mao, Yuxi Qian, Junjie Ye, Hang Zhao, and Yue
385 Wang. Gpt-driver: Learning to drive with gpt. *arXiv preprint
386 arXiv:2310.01415*, 2023. 2
- 387 [14] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley,
388 and Jianfeng Gao. Instruction Tuning with GPT-4, 2023.
389 *arXiv:2304.03277* [cs]. 1
- 390 [15] Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen,
391 Hanxue Zhang, Chengen Xie, Jens Beißenwenger, Ping Luo,
392 Andreas Geiger, and Hongyang Li. DriveLM: Driving with
393 Graph Visual Question Answering, 2025. *arXiv:2312.14150*
394 [cs]. 1
- 395 [16] Ran Tian, Boyi Li, Xinshuo Weng, Yuxiao Chen, Edward
396 Schmerling, Yue Wang, Boris Ivanovic, and Marco Pavone.
397 Tokenize the world into object-level knowledge to address
398 long-tail events in autonomous driving. *arXiv preprint
399 arXiv:2407.00959*, 2024. 2
- 400 [17] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang
401 Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng
402 Lang, and Hang Zhao. DriveVLM: The Convergence of
403 Autonomous Driving and Large Vision-Language Models,
404 2024. *arXiv:2402.12289* [cs]. 1
- 405 [18] Joseph Raj Vishal, Divesh Basina, Aarya Choudhary, and
406 Bharatesh Chakravarthi. Eyes on the Road: State-of-the-Art
407 Video Question Answering Models Assessment for Traffic
408 Monitoring Tasks, 2024. *arXiv:2412.01132* [cs]. 1
- 409 [19] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu,
410 Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue
411 Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Ji-
412 awei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang
413 Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong,
414

- 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 Letters*,
 423 2024. 2
- 424 [21] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying
 425 Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Dar-
 426 rrell. 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 SEQUENCIAL 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.