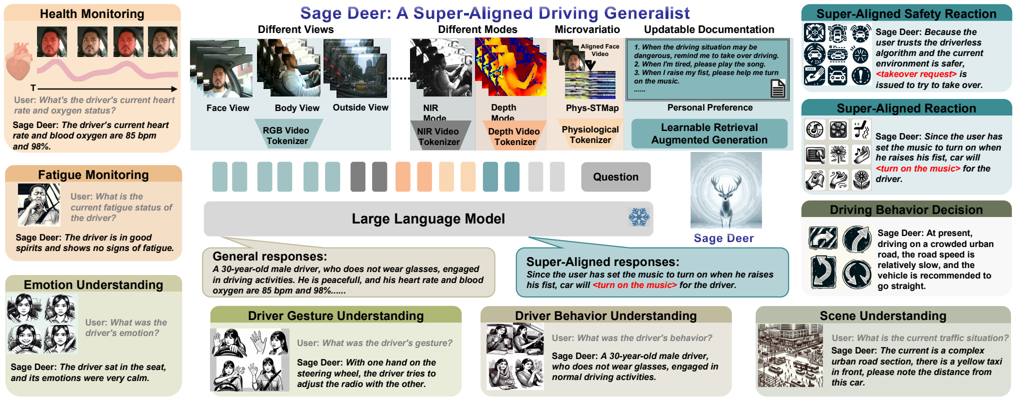
A SUPER-ALIGNED DRIVING GENERALIST IN YOUR COCKPIT

Anonymous authors
Paper under double-blind review



User: *What's the driver's current heart rate and oxygen status?*
Sage Deer: *The driver's current heart rate and blood oxygen are 85 bpm and 98%.*

User: *What is the current fatigue status of the driver?*
Sage Deer: *The driver is in good spirits and shows no signs of fatigue.*

User: *What was the driver's emotion?*
Sage Deer: *The driver sat in the seat, and its emotions were very calm.*

1. When the driving situation may be dangerous, remind me to take over driving.
2. When I'm tired, please play the song.
3. When I raise my fist, please help me turn on the music.
......

General responses: *A 30-year-old male driver, who does not wear glasses, engaged in driving activities. He is peacefull, and his heart rate and blood oxygen are 85 bpm and 98%......*

Super-Aligned responses: *Since the user has set the music to turn on when he raises his fist, car will <turn on the music> for the driver.*

User: *What was the driver's gesture?*
Sage Deer: *With one hand on the steering wheel, the driver tries to adjust the radio with the other.*

User: *What was the driver's behavior?*
Sage Deer: *A 30-year-old male driver, who does not wear glasses, engaged in normal driving activities.*

User: *What is the current traffic situation?*
Sage Deer: *The current is a complex urban road section, there is a yellow taxi in front, please note the distance from this car.*

Sage Deer: *Because the user trusts the driverless algorithm and the current environment is safer, <takeover request> is issued to try to take over.*

Sage Deer: *Since the user has set the music to turn on when he raises his fist, car will <turn on the music> for the driver.*

Sage Deer: At present, driving on a crowded urban road, the road speed is relatively slow, and the vehicle is recommended to go straight.

Figure 1: Sage Deer uses the pre-trained video encoders (visual tokenizers) to tokenize the different modes and views of the video, especially the physiological encoder used to extract the physiological signals of face video. Using tokenized visual embedding, large language models can provide general responses for physiological indicators, emotional states, gestures, human behavior, and scene understanding. It is worth emphasizing that the individual needs of the user can be edited in a single document. We then query this information with learnable retrieval augmentation generation, generating results super aligned with user preferences.

ABSTRACT

The intelligent driving cockpit, an important part of intelligent driving, needs to match different users' comfort, interaction, and safety needs. This paper aims to build a **s**uper-**a**ligned and **ge**neralist **dr**iving agent, **sage deer**. Sage Deer achieves two highlights: (1) Super alignment: It achieves different reactions according to different people's preferences and biases. (2) Generalist: It can understand the user's physiological indicators, facial emotions, hand movements, body movements, driving scenarios, and behavioral decisions. (3) Multimodal: He can understand RGB, NIR, and depth video to build more robust perception, understanding, and reasoning. To achieve the above requirements, we design retrieval-enhanced multimodal frameworks. We collected multiple data sets and built a large-scale benchmark. This benchmark measures the deer's perceptual decision-making ability and the super alignment's accuracy.

1 INTRODUCTION

Owing to advancements in artificial intelligence and high performance computing hardware, intelligent connected vehicles have garnered significant attention from both academia and governmental bodies due to their profound potential to revolutionize future transportation paradigms Li et al. (2023b); Teng et al. (2023). By integrating advanced sensors, artificial intelligence algorithms, and technologies, s possess the capability to perceive their environment, make informed decisions, and execute control, which will significantly enhance vehicle safety, optimize traffic flow, and deliver a comfortable and convenient driving experience.

The intelligent driving cockpit serves as the interface for human interaction with s, offering a seamless and user-focused experience through the provision of real-time data, personalized settings, and intuitive controls Li et al. (2023b); Yang et al. (2022). This advanced system enhances communication between the driver, passengers, and the vehicle, increases safety via comprehensive monitoring systems, and improves comfort by adapting to individual preferences and varying driving conditions.

However, the pursuit of a universal intelligent driving cockpit that encapsulates both general principles and individual-specific requirements is inherently complex. Individuals possess distinct driving habits and needs, necessitating the development of an agent that aligns with their preferences. However, it is impractical to train a unique model for each individual user. Existing research Cui et al. (2023b;a); Li et al. (2023c; 2022); Liu & Zhang (2020); Tao et al. (2024) on intelligent driving cockpits predominantly emphasizes rigid decision-making and control by integrating external environmental perception data and in-vehicle occupant status, which usually overlooks the individual preferences and limits the adaptability of cockpit system.

Recent advancements in multi-modal large models have demonstrated their capabilities in understanding and reasoning with multi-modal inputs, including videos and natural language. Sima et al. Sima et al. (2023) proposed the utilization of large language models to perform end-to-end autonomous driving through a visual question answering manner. While Wang et al. Wang et al. (2023) proposed DriveMLM, which can perform close-loop autonomous driving in realistic simulators by mitigate the gap between language decisions and vehicle control commands. As for in-vehicle situations, multi-modal large models are also applied to driver health monitoring Hecht et al. (2018), driving decision individuation with simple strategy Cui et al. (2023b). A robust and personalized intelligent driving cockpit can incorporate driving scenario understanding, language interaction, user behavior and sentiment analysis, processing of personalized driver needs, and intelligent decision-making. However, the above mentioned research has yet to achieve comprehensive functionality and full-modal information processing and reasoning. Thus, there remains a significant gap between current capabilities and the envisioned intelligent driving cockpit.

To bridge the gap between current approaches in providing a personalized driving experience and comprehensive omni-modal information processing, we propose a **s**uper-**a**ligned and **ge**neralist **dr**iving agent, **sage deer**. We present a retrieval-augmented framework designed to develop a super-aligned and generalist agent for smart cockpits. This framework is characterized by three core strengths: comprehensive tokenization of multi-mode and multi-view videos, robust generalist understanding, and super-aligned responses. To effectively integrate diverse visual and physiological data, we tokenize inputs from multiple sensing modalities—including RGB, near-infrared (NIR), and depth cameras—using a pre-trained ResNet18 for feature extraction, followed by mapping to a language space with a two-layer linear layer. Multi-view perspectives are similarly tokenized with distinct perspective markers to ensure comprehensive scene interpretation. Additionally, physiological information is processed through spatio-temporal representations and encoded into tokens, enabling the model to monitor driver states accurately. The framework employs a retrieval-augmented generation mechanism that leverages an updatable document containing both fundamental and personalized information, enhancing the large language model's (LLM) ability to generate contextually relevant responses without extensive fine-tuning. Expert knowledge fusion further integrates domain-specific insights, allowing the system to dynamically adapt based on real-time physiological and environmental data. This holistic approach ensures high-quality scene understanding, precise physiological monitoring, and personalized, context-aware interactions, significantly improving the alignment and generalization capabilities of agents within complex cockpit environments.

Contribution

1. Established a multi-view and multi-modal evaluation protocol for the intelligent cockpit, involving unified evaluation of the driver's physiology, emotion, behavior, driving scene understanding, and decision-making.

2. An intelligent driving cockpit super alignment evaluation protocol involving generalization ability for different needs was established.

3. Designed a set of algorithms that align human preferences without fine-tuning.

Table 1: Comparison of public driving cockpit perception datasets. "Naturalistic" refers to people's natural driving behavior, while "Induced" refers to people's intervention and design of scenes. Reaction is the response given by the intelligent cockpit according to different user needs. Question Answering refers to the verbal inquiry interaction between the user and the cockpit.

Dataset	Views	Naturalistic	Induced	Multimodal	Physiological Indicator	Behavior	Emotion	Traffic Context	Vehicle Condition	Reaction	Question Answering
SEU Zhao et al. (2012)	1	–	✓	–	–	✓	–	–	–	–	–
Tran Tran et al. (2018)	1	–	✓	–	–	✓	–	–	–	–	–
Zhang Zhang et al. (2020)	2	–	✓	✓	–	✓	–	–	–	–	–
StateFarm sta (2016)	1	–	✓	–	–	✓	–	–	–	–	–
AUC-DD Eraqi et al. (2019)	1	✓	–	–	–	✓	–	–	–	–	–
LoLi Saad et al. (2020)	1	✓	–	✓	–	✓	–	–	–	–	–
Brain4Cars Jain et al. (2016)	2	✓	–	–	–	✓	–	–	–	–	–
Drive&Act Martin et al. (2019)	6	–	✓	✓	–	✓	–	–	–	–	–
DMD Ortega et al. (2020)	3	–	✓	✓	–	✓	–	–	–	–	–
DAD Kopuklu et al. (2021)	2	–	✓	✓	–	✓	–	–	–	–	–
DriPE Guesdon et al. (2021)	1	✓	–	–	–	–	–	–	–	–	–
LBW Kasahara et al. (2022)	2	✓	–	–	–	–	–	–	–	–	–
MDAD Jegham et al. (2019)	2	✓	–	✓	–	✓	–	–	–	–	–
3MDAD Jegham et al. (2020)	2	✓	–	✓	–	✓	–	–	–	–	–
DEFE Li et al. (2021a)	1	–	✓	–	–	–	✓	–	–	–	–
DEFE+ Li et al. (2021b)	1	–	✓	✓	–	–	✓	–	–	–	–
Du Du et al. (2020)	1	–	✓	✓	–	–	✓	–	–	–	–
KMU-FED Jeong & Ko (2018)	1	✓	–	–	–	–	✓	–	–	–	–
MDCS Oh et al. (2022)	2	✓	–	✓	–	–	✓	–	–	–	–
AIDE Yang et al. (2023a)	4	✓	–	–	–	✓	✓	✓	✓	–	–
Sage Deer	4	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

2 Related Work

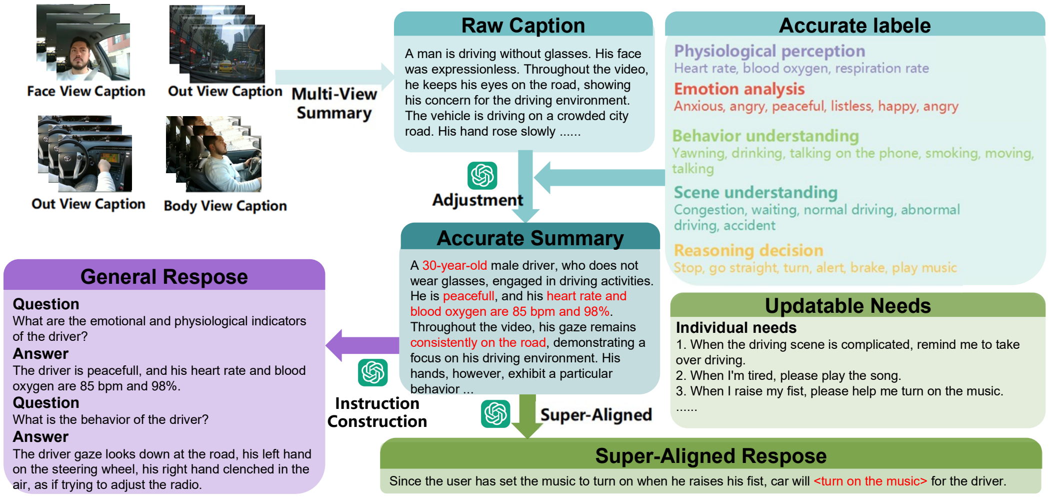
Multi-modality Large Model in Driving Multi-modality large models are widely used in autonomous driving, such as automatic planning and control Mao et al. (2023); Cui et al. (2023a), perception Wang et al. (2020), and driver health monitoring Hecht et al. (2018) among others. By aggregating multi-modal information (e.g., vision, speech, point cloud, etc.) Yang et al. (2023b), multi-modal large models can more effectively help autonomous driving systems perceive the scene inside and outside the vehicle, and make more intelligent decisions.

In the vehicle's user-facing interface, the driver's emotions, health, posture, and actions can help the autonomous driving system interact dynamically and adapt to various driving styles. Cui et al. Cui et al. (2023b;a) first proposed to precisely enhance the decision-making process of autonomous driving by perceiving multi-modal information inside and outside the vehicle. Although these methods can perceive multi-modal information, they adopt a single strategy for different users and cannot make adaptive decisions for different users. Subsequent improvements Cui et al. (2023c); Yang et al. (2024) attempt to promote fine-grained understanding of user language and enhance the control performance of the vehicle system, but these methods overlook the impact of hidden emotions, health, and action information on vehicle safety and efficiency.

In this paper, we super-align various modal information of users, which not only includes the user's own interaction, but also perceives the user's health, emotions, posture, behavior, and other information, thereby enhancing the intelligence of the vehicle system.

Retrieval-Augmented Generation Retrieval-Augmented Generation (RAG) enhances the accuracy of knowledge-intensive tasks by retrieving relevant information from an external knowledge base, which can be continuously updated Gao et al. (2023). Specifically, RAG first collects external information and saves old user queries Ma et al. (2023), organizes the knowledge, aligns it, and stores it in a repository provided by the user. Then, an efficient index is built to retrieve image data. Afterward, RAG retrieves internal information based on the prompts provided by the user, which mainly relies on priority ranking and retrieval. Finally, RAG merges the proposed query and selected information inputs it into a large model and returns a more accurate response. Based on this, RAG can help answer questions that vary according to specific needs while also aggregating the dialogue history scenario, enabling the model to remember customized information effectively.

In this paper, we use RAG to store customized information of users, including their driving habits and behaviors, which helps build a more human-centered vehicle driving agent, and can use customized requirements of specific users.



Raw Caption

A man is driving without glasses. His face was expressionless. Throughout the video, he keeps his eyes on the road, showing his concern for the driving environment. The vehicle is driving on a crowded city road. His hand rose slowly ......

Accurate labele

Physiological perception
Heart rate, blood oxygen, respiration rate
Emotion analysis
Anxious, angry, peaceful, listless, happy, angry
Behavior understanding
Yawning, drinking, talking on the phone, smoking, moving, talking
Scene understanding
Congestion, waiting, normal driving, abnormal driving, accident
Reasoning decision
Stop, go straight, turn, alert, brake, play music

Accurate Summary

A 30-year-old male driver, who does not wear glasses, engaged in driving activities. He is peacefull, and his heart rate and blood oxygen are 85 bpm and 98%. Throughout the video, his gaze remains consistently on the road, demonstrating a focus on his driving environment. His hands, however, exhibit a particular behavior ...

General Respose

Question
What are the emotional and physiological indicators of the driver?
Answer
The driver is peacefull, and his heart rate and blood oxygen are 85 bpm and 98%.
Question
What is the behavior of the driver?
Answer
The driver gaze looks down at the road, his left hand on the steering wheel, his right hand clenched in the air, as if trying to adjust the radio.

Updatable Needs

Individual needs
1. When the driving scene is complicated, remind me to take over driving.
2. When I'm tired, please play the song.
3. When I raise my fist, please help me turn on the music.
......

Super-Aligned Respose

Since the user has set the music to turn on when he raises his fist, car will <turn on the music> for the driver.

Figure 2: (a) We use existent multimodal model tools to generate captions for the videos likes (Li et al., 2023a; Zhang et al., 2023a). (b) Then, we set a reasonable prompt to merge the information from different videos. (c) We took advantage of the existing tags (including physiological indicators, emotional indicators, action indicators, behavioral indicators, scene understanding, and reasoning decision-making) to correct and supplement the captain. (d) Next, we use GPT4 (Achiam et al., 2023) as an assistant to build question-answering pairs for different tasks (including physiological indicators, emotion, behavior, and so on.). (e) We design multiple user preferences, and GPT4 responds to the current scenario based on user preferences.

3 Data Curating and Beachmarking

3.1 Data collection

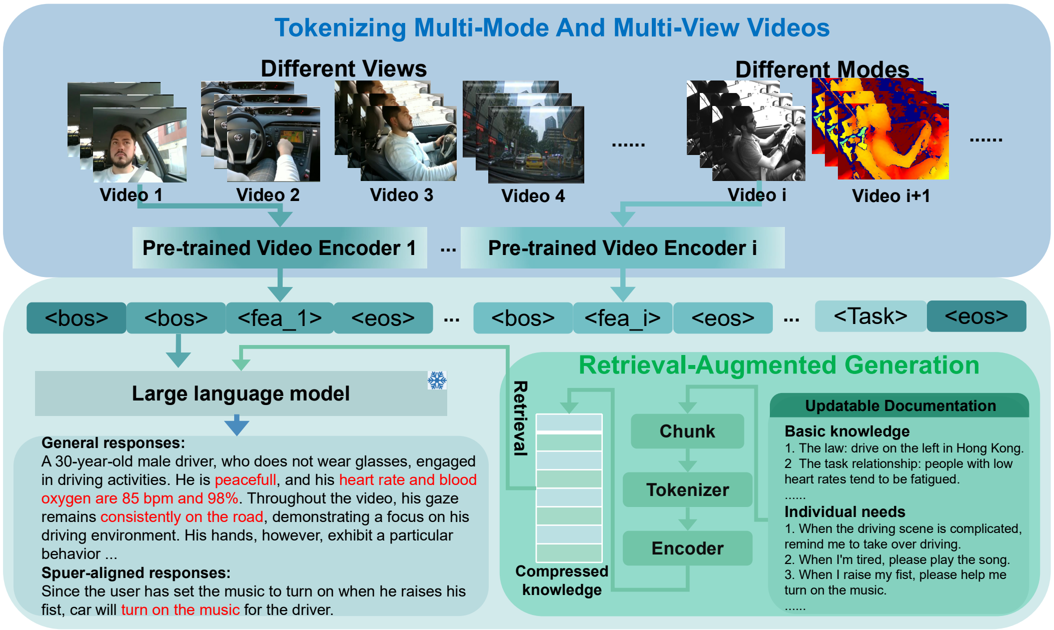
Our goal is to create a driving generalist (physiological estimation, emotional estimation, gesture estimation, body motion estimation, driving behavior estimation, driving behavior detection, and driving decision-making) in the intelligent cockpit. We selected three of the most recent multi-view multi-task driving datasets AIDE and DMD. In addition, we use the latest contactless physiological measurement technology to monitor the user's health, which can estimate heart rate and blood oxygen only with a camera. We collected 5 datasets (VIPL-HR, V4V, PURE, BUAA-rPPG and UBFC).

3.2 General Instruction Construction

To construct natural language descriptions for sage deer, we use existent multimodal model tools (InternVideo (Wang et al., 2022), Tag2Text (Huang et al., 2023), or GRiT (Wu et al., 2022)) to generate captions for the frame sampled at equal intervals automatically. Likes (Li et al., 2023a; Zhang et al., 2023a), we set the reasonable prompt to merge the information of different frames. Existing image caption methods are often inaccurate or insufficiently annotated for an intelligent cockpit system. So, we took advantage of the existing tags (including physiological indicators, emotional indicators, action indicators, behavioral indicators, scene understanding, and reasoning decision-making) in the dataset to correct and supplement the captain. Next, we use GPT4 (Achiam et al., 2023) as an assistant to build question-answering pairs for different tasks. We used GPT4 to summarize, fuse, and correct these captains. See supplementary materials for details.

3.3 Super-Aligned Reaction

Users have different needs for driving cockpit, especially interactivity and trustworthiness. Interactivity means creating gestures, emotions, and body movements and tailoring cockpit feedback to user needs. Trustworthiness refers to personalized warning feedback on user fatigue, bad mood, and bad behavior. For this purpose, we use GPT4 to design the needs of different users and give specific feedback according to the needs of users and the information of the current scenario.



General responses:
A 30-year-old male driver, who does not wear glasses, engaged in driving activities. He is peacefull, and his heart rate and blood oxygen are 85 bpm and 98%. Throughout the video, his gaze remains consistently on the road, demonstrating a focus on his driving environment. His hands, however, exhibit a particular behavior ...

Spuer-aligned responses:
Since the user has set the music to turn on when he raises his fist, car will turn on the music for the driver.

Basic knowledge
1. The law: drive on the left in Hong Kong.
2 The task relationship: people with low heart rates tend to be fatigued.
......

Individual needs
1. When the driving scene is complicated, remind me to take over driving.
2. When I'm tired, please play the song.
3. When I raise my fist, please help me turn on the music.
......

Figure 3: Overview of our framework. RGB, depth and NIR videos of different modes are extracted by a pre-trained encoder and converted into tokens. These tokens concatenate problems and retrieve enhanced linguistic features. Where the search augmentation generated to language features are now features that will be generated by the interchangeable document, via chunk, tokenizer, and further encoder. Ultimately, the LLM will respond based on the updatable document content and video content.

4 METHOD

We build a retrieval-augmented framework to build a super-aligned and generalist agent in the cockpit. This framework has three strengths: (1) Any mode and view video tokenization. (2) Generalists understand. (3) Super-aligned reactions.

4.1 TOKENIZING MULTI-MODE AND MULTI-VIEW VIDEOS

For driving cockpit, leveraging multiple modes and perspectives can significantly enhance the model's ability to interpret complex scenes, especially under challenging conditions such as poor weather or low-light environments. This subsection outlines our approach to tokenizing multi-mode and multi-view videos, ensuring that the model effectively integrates diverse visual information and physiological indicators.

Tokenizing Multi-Model. To robustly monitor user and scene information under adverse conditions, we incorporate multiple sensing modalities, including RGB, near-infrared (NIR), and depth cameras. We uniformly use the ImageNet pre-trained ResNet18 as each frame rate feature extractor, and then we cat all the frames together to get the video features. Finally, a simple two-layer linear layer is mapped to the language space$em_{face} \in C \times L$. C is the number of channels characteristic of the language model, and L is a hyperparameter representing how many tokens are used to present the video features. In particular, to better understand the physiological indicators of faces in the video. To allow LLM to distinguish between different modes, we added identifiers for visual features. For example, add corresponding start and end symbols to the front and back of a face RGB video embedding $em_{rgb} \in C \times L, < \ RGB\ bos > \ em_{rgb} \ < \ RGB\ cos >$. Other videos are processed the same way and then fed into the LLM.

Tokenizing Multi-View. Input of multiple perspective information is necessary, such as viewing information around the entire vehicle or monitoring the scene and driver at the same time. Similarly to multi-model, we uniformly use the ImageNet pre-trained ResNet18 as the feature extractor for each frame rate. This is followed by mapping a simple two-layer linear layer to the language space. Then, we also use perspective markers to mark different perspectives, for example, $em_{front} \in C \times L$, $< Front\ RGB\ bos >\ em_{front}\ < Front\ RGB\ cos >$.

Tokenizing Physiological Information. Through a contactless video of a person's face, we can obtain physiological indicators such as heart rate, blood oxygen and respiration. But directly feeding a long video into our model is very low effect. Following (Lu et al., 2023), we first convert the video into a spatio-temporal representation (STMap) and then extract the features with ResNet18 and convert them into tokens $em_{phys} \in C \times L$ through two linear layers. Similarly, we've added identifiers, $< Phys\ bos >\ em_{phys}\ < Phys\ cos >$

Final representation. In addition, the user's questions are concated with vision tokens and sent to the LLM. The losses and training strategy of our visual instruction tuning are the same as those of MLLM (Li et al., 2023a; Zhang et al., 2023a; Muhammad Maaz & Khan, 2023).

4.2 Retrieval-Augmented Generation

The training of large language models is very expensive, and it is easy to lose the ability to generalize in a small training data. User needs for a smart cockpit are often very different, but we can't fine-tune the LLM to each user's specific preferences. To do this, we have adopted a retrieval-augmented generation framework ash shown in Fig. 3.

Specifically, we built an updatable document. This document includes some basics and personal requirements. Then we chunk the document, here we use a very concise way to divide the document by sentences. Each sentence is then tokenized, and then the sentence is the same length by filling in the tokens. A four-layer convolutional encoder is then used to compress the literal token into a 1/8-length feature. Then, we calculate the similarity between visual features and text features, and only retain the first N features with high similarity. Splicing into visual features. Then, we calculate the similarity between visual features and text features, and only retain the first N features with high similarity. These retrieved temporal features are spliced onto visual features.

4.3 Expert Knowledge Fusion

Reasonable knowledge and carefully crafted prompts can significantly enhance the potential of Large Language Models (LLMs), enabling them to generate more nuanced and context-aware outputs. In the context of the smart cockpit, there exists a strong and intricate relationship between various factors such as physiological indicators, emotions, behaviors, and the vehicle itself. Understanding these correlations is crucial for optimizing driver assistance systems, improving safety, and providing a more personalized driving experience. To this end, we conducted a systematic investigation of task correlations in collaboration with vehicle human factors experts, enabling us to explore these dependencies in greater detail.

The Relationship Between Physiological State and Emotion: Classical studies in psychophysiology (e.g., James-Lange, Schachter-Singer) have long established that physiological states, such as heart rate variability, skin conductance, and muscle tension, are closely linked to emotional responses. In the context of smart cockpits, real-time monitoring of physiological signals can provide valuable insights into the emotional state of the driver. For instance, elevated heart rate and increased skin conductance may be indicative of stress or anxiety, particularly in challenging driving scenarios. Recognizing this connection allows the system to adapt accordingly, offering calming feedback or assistance in real-time.

In addition, we investigated prior knowledge of other tasks, including: The Relationship Between Physiological State and Behavior,The Relationship Between Physiological State and Behavior, The Relationship Between Physiological State and Behavior. These relationships are not isolated but rather dynamically interact in real-world driving scenarios. By leveraging real-time data on cabin conditions — including environmental factors, driver behavior, and physiological/emotional states — the smart cockpit can generate a wider range of more accurate and contextually appropriate responses. This dynamic adaptation not only enhances the comfort and safety of the driver but also improves the overall driving experience. For example, in high-stress situations, the system may offer enhanced alerts or autonomous driving support, while in calm conditions, it might prioritize fuel efficiency or offer entertainment options.

Table 2: Baseline compared the performance of our approach on the super alignment protocol. Baseline is to splice user requirements directly into visual features and feed them into the LLM, taking advantage of the long-range modeling capabilities of the LLM itself.

Method	**AIDE**		**DMD**	
	BLEU	SPICE	BLEU	SPICE
Baseline	0.184	0.301	0.191	0.296
Ours	**0.214**	**0.315**	**0.204**	**0.305**

By embracing the interconnected nature of these factors, intelligent systems in smart cockpits can better understand the human-driver interaction, anticipating needs and providing tailored responses that enhance both safety and the overall driving experience. Importantly, this approach underscores the potential of LLMs when incorporated with domain-specific knowledge, allowing them to generate more context-sensitive and accurate outputs in real time. Integrating such knowledge-driven, prompt-based systems into smart cockpits represents a significant leap in creating more intuitive, responsive, and adaptive vehicle environments.

5 EXPERIMENTS

5.1 DATASETS.

We found two recent data sets of very comprehensive annotations for assisted driving. The DMD is a driver monitoring dataset, an extensive dataset that includes real and simulated driving scenarios: distraction, eye distribution, drowsiness, handwheel interaction, and contextual data, in 41 hours of RGB, depth, and infrared video from 3 cameras, capturing the faces, bodies, and hands of 37 drivers. AIDE proposes an assisted driving awareness dataset that takes into account contextual information both inside and outside the vehicle in a natural scenario. AIDE enables overall driver monitoring through multi-perspective Settings of the driver and the scene, multi-modal annotations of the face, body, posture, and gestures, and four practical task designs for driver understanding. we collect five rPPG face video datasets (VIPL-HR Niu et al. (2019), PURE Stricker et al. (2014), UBFC-rPPG Bobbia et al. (2019), V4V Revanur et al. (2021), and BUAA-MIHR Xi et al. (2020)), mostly with subjects remaining still, and some with head movements.

5.2 TRAINING DETAILS.

We tain our model on A6000 for 2 epochs. The learning rate, and weight decay are set to 0.001, and 0.02, respectively. The maximum sentence length is set to 64. That is, if the sentence is too long, excess parts will be discarded, and if the sentence is too short, 0 will be filled to make the length uniform. All image vision encoders use the pre-trained resnet 18 on imagenet as the pre-trained model, and these vision encoders are fine-tuned. For the feature extraction of physiological signals, we use the pre-trained model of NEST-rPPG Lu et al. (2023) and do not fine-tune it.

5.3 BASELINES AND EVALUATION METRICS

We compare our proposed framework's anomaly understanding performance with SOTA video understanding baselines. We select five baselines: Video-ChatGPT (Muhammad Maaz & Khan, 2023), VideoChat (Li et al., 2023a), Video-LLaMA (Zhang et al., 2023a), LLaMA-Adapter (Zhang et al., 2023b), and Video-LLaVA (Lin et al., 2023). Our comparison aims to determine whether these baselines can fully understand and interpret video anomalies.

To accurately evaluate our model's performance, we adopt BLEU Bilingual Evaluation Understudy(BLEU) and SPICE (Papineni et al., 2002) to measure word overlap between the model-generated text and the ground truth. This approach enables us to objectively assess the similarity and consider various levels of granularity at the text level, thus clearly indicating how well the model understands and describes anomalies.

Table 3: Generalist capabilities on AIDE data sets.

Method	**Emotion**		**Behavior**		**Scene**		**Condition**	
	BLEU	SPICE	BLEU	SPICE	BLEU	SPICE	BLEU	SPICE
Video-ChatGPT	0.200	0.280	0.170	0.205	0.190	0.330	0.195	0.320
VideoChat	0.205	0.285	0.175	0.210	0.195	0.335	0.200	0.325
Video-Llama	0.217	0.312	0.183	0.223	0.207	0.352	0.211	0.337
Llama-Adapter	0.215	0.310	0.190	0.225	0.210	0.340	0.189	0.321
Video-LLaVA	0.202	0.290	0.172	0.205	0.190	0.330	0.195	0.320
Ours	**0.232**	**0.331**	**0.194**	**0.242**	**0.225**	**0.369**	**0.223**	**0.360**

Table 4: Generalist capabilities on DMD data sets.

Method	**Action**		**Gaze**		**Hand**	
	BLEU	SPICE	BLEU	SPICE	BLEU	SPICE
Video-ChatGPT	0.175	0.250	0.150	0.185	0.160	0.310
VideoChat	0.190	0.260	0.160	0.185	0.160	0.310
Video-Llama	0.205	0.275	0.170	0.205	0.180	0.325
Llama-Adapter	0.202	0.265	0.170	0.201	0.175	0.320
Video-LLaVA	0.195	0.295	0.155	0.220	0.170	0.315
Ours	**0.235**	**0.315**	**0.195**	**0.230**	**0.210**	**0.340**

5.4 Generalist Performance

Our model can estimate the driver's emotion, physiological indicators, gaze, physical behavior, hand behavior, driving scene and vehicle state. In order to more clearly evaluate the ability of the model in different subtasks, we conducted systematic evaluation on two multi-task datasets, fatigue and physiological indicators. First, we used the default dataset partitioning on the AIDE dataset to train and test the results shown in Table 3. As you can see from the table, our algorithm has improved tremendously. This is mainly because, using expert knowledge, LLM can better integrate the connections between tasks.

Further, we verify our algorithm on DMD data set. As shown in Table 4, our algorithm achieves excellent performance on all tasks. All these results prove the effectiveness of our algorithm. Of course, our algorithm is very concise, only through RAG to retrieve the relevant expert knowledge to improve the accuracy of the reply. Further, we give an example of our model's dialogue to illustrate the generalist power of our model 4.

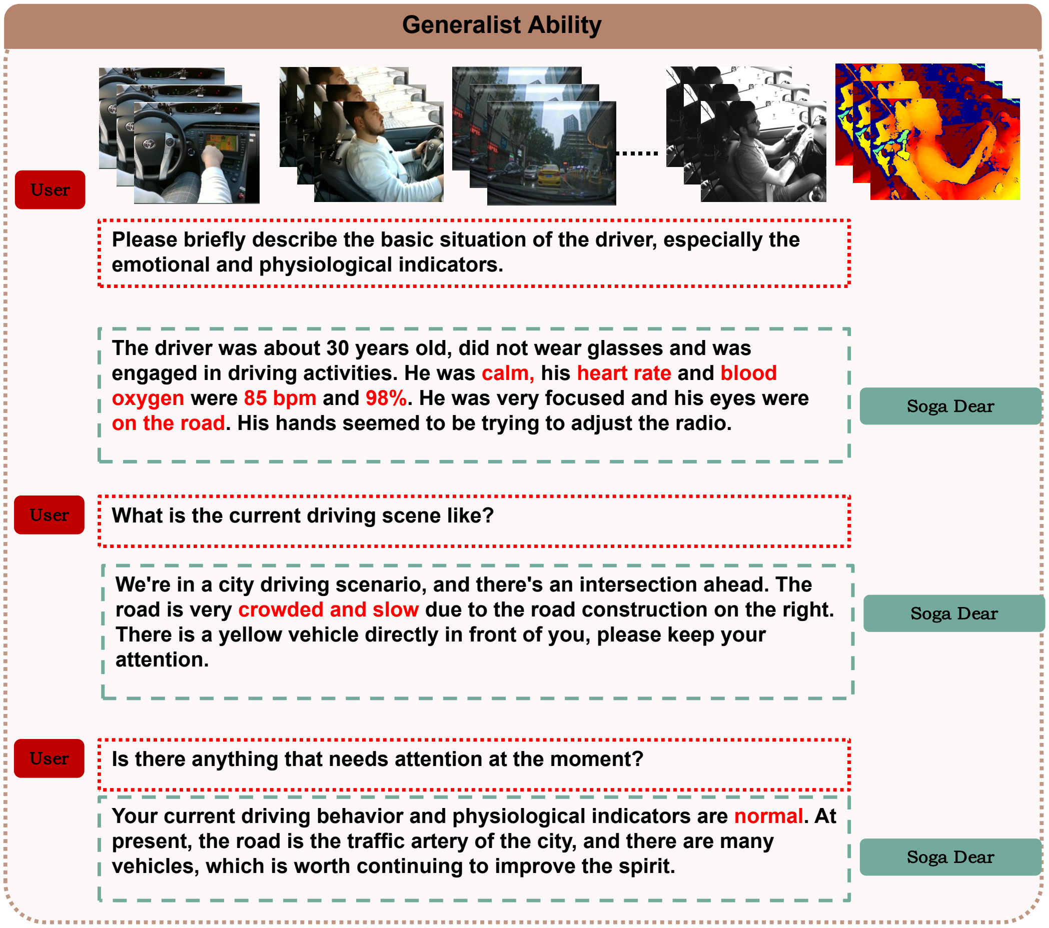
5.5 Super-Aligned Performance

Different users have different needs for driving a smart cockpit, which requires us to super-align our users' biases. One of the simplest ways is to put all the user's preferences in the form of text at the top of the LLM input, and then use the LLM's own long-distance modeling capabilities to bias it. This most straightforward approach is considered our Baseline. Our approach is to complete the super alignment with a learned RAG. The performance of these two methods is shown in Table 2.

As shown in Table 2, our approach has improved dramatically. This shows that LLM itself has a limited ability to understand long texts, and it is difficult to make different responses to the results according to its visual features and prior knowledge. This explains the need for our RAG framework.

6 Conclusion

In this study, we successfully curated and benchmarked comprehensive data to develop Sage Deer, a super-aligned and generalist driving agent for intelligent cockpits. Our data collection strategy integrated three state-of-the-art multi-view and multi-task driving datasets—AIDE, DMD, VIPL-HR,



Please briefly describe the basic situation of the driver, especially the emotional and physiological indicators.

The driver was about 30 years old, did not wear glasses and was engaged in driving activities. He was calm, his heart rate and blood oxygen were 85 bpm and 98%. He was very focused and his eyes were on the road. His hands seemed to be trying to adjust the radio.

What is the current driving scene like?

We're in a city driving scenario, and there's an intersection ahead. The road is very crowded and slow due to the road construction on the right. There is a yellow vehicle directly in front of you, please keep your attention.

Is there anything that needs attention at the moment?

Your current driving behavior and physiological indicators are normal. At present, the road is the traffic artery of the city, and there are many vehicles, which is worth continuing to improve the spirit.

Figure 4: Generalist ability. Soga deer can reply to different tasks and has the ability to reply to an open domain.

V4V, PURE, BUAA-rPPG, and UBFC—alongside cutting-edge contactless physiological measurement technologies to monitor vital health indicators such as heart rate and blood oxygen levels using only camera inputs. To construct robust natural language instructions, we employed advanced multimodal models (InternVideo, Tag2Text, GRiT) for automatic video captioning, which were then refined and enhanced with existing tags and supplemented through GPT-4-generated question-answer pairs tailored to various driving-related tasks. Additionally, we addressed diverse user needs by leveraging GPT-4 to design personalized interactions, ensuring that the intelligent cockpit could provide tailored feedback based on individual preferences and real-time scenarios. Our benchmarking framework evaluates the agent's ability to interpret physiological data, emotional states, gestures, body movements, driving behaviors, and decision-making processes across multiple modalities. The integration of retrieval-augmented generation and expert knowledge fusion within our framework ensures that Sage Deer delivers accurate, context-aware, and personalized responses, significantly advancing the capabilities of intelligent driving cockpits in enhancing driver comfort, safety, and interaction.

REFERENCES

State farm distracted driver detection, 2016. `https://www.kaggle.com/c/state-farm-distracted-driver-detection`.

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.

Serge Bobbia, Richard Macwan, Yannick Benezeth, Alamin Mansouri, and Julien Dubois. Unsupervised skin tissue segmentation for remote photoplethysmography. *Pattern Recognition Letters*, 124:82–90, 2019.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 902–909, 2023a. URL `https://api.semanticscholar.org/CorpusID:262054629`.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. Receive, reason, and react: Drive as you say with large language models in autonomous vehicles. *ArXiv*, abs/2310.08034, 2023b. URL `https://api.semanticscholar.org/CorpusID:263908840`.

Can Cui, Zichong Yang, Yupeng Zhou, Yunsheng Ma, Juanwu Lu, and Ziran Wang. Personalized autonomous driving with large language models: Field experiments. 2023c. URL `https://api.semanticscholar.org/CorpusID:266335383`.

Guanglong Du, Zhiyao Wang, Boyu Gao, Shahid Mumtaz, Khamael M Abualnaja, and Cuifeng Du. A convolution bidirectional long short-term memory neural network for driver emotion recognition. *IEEE Transactions on Intelligent Transportation Systems*, 22(7):4570–4578, 2020.

Hesham M Eraqi, Yehya Abouelnaga, Mohamed H Saad, and Mohamed N Moustafa. Driver distraction identification with an ensemble of convolutional neural networks. *Journal of Advanced Transportation*, 2019, 2019.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997, 2023. URL `https://api.semanticscholar.org/CorpusID:266359151`.

Romain Guesdon, Carlos Crispim-Junior, and Laure Tougne. Dripe: A dataset for human pose estimation in real-world driving settings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2865–2874, 2021.

Tobias Hecht, Anna Feldhütter, Jonas Radlmayr, Yasuhiko Nakano, Yoshikuni Miki, Corbinian Henle, and Klaus Bengler. A review of driver state monitoring systems in the context of automated driving. *Advances in Intelligent Systems and Computing*, 2018. URL `https://api.semanticscholar.org/CorpusID:56790391`.

Xinyu Huang, Youcai Zhang, Jinyu Ma, Weiwei Tian, Rui Feng, Yuejie Zhang, Yaqian Li, Yandong Guo, and Lei Zhang. Tag2text: Guiding vision-language model via image tagging. *arXiv preprint arXiv:2303.05657*, 2023.

Ashesh Jain, Hema S Koppula, Shane Soh, Bharad Raghavan, Avi Singh, and Ashutosh Saxena. Brain4cars: Car that knows before you do via sensory-fusion deep learning architecture. *arXiv preprint arXiv:1601.00740*, 2016.

Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. Mdad: A multimodal and multiview in-vehicle driver action dataset. In *International Conference on Computer Analysis of Images and Patterns*, pp. 518–529, 2019.

Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. A novel public dataset for multimodal multiview and multispectral driver distraction analysis: 3mdad. *Signal Processing: Image Communication*, 88:115960, 2020.

Mira Jeong and Byoung Chul Ko. Driver's facial expression recognition in real-time for safe driving. *Sensors*, 18(12):4270, 2018.

Isaac Kasahara, Simon Stent, and Hyun Soo Park. Look both ways: Self-supervising driver gaze estimation and road scene saliency. In *European Conference on Computer Vision (ECCV)*, pp. 126–142, 2022.

Okan Kopuklu, Jiapeng Zheng, Hang Xu, and Gerhard Rigoll. Driver anomaly detection: A dataset and contrastive learning approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 91–100, 2021.

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

Wenbo Li, Yaodong Cui, Yintao Ma, Xingxin Chen, Guofa Li, Guanzhong Zeng, Gang Guo, and Dongpu Cao. A spontaneous driver emotion facial expression (defe) dataset for intelligent vehicles: Emotions triggered by video-audio clips in driving scenarios. *IEEE Transactions on Affective Computing*, 2021a.

Wenbo Li, Guanzhong Zeng, Juncheng Zhang, Yan Xu, Yang Xing, Rui Zhou, Gang Guo, Yu Shen, Dongpu Cao, and Fei-Yue Wang. Cogemonet: A cognitive-feature-augmented driver emotion recognition model for smart cockpit. *IEEE Transactions on Computational Social Systems*, 9(3): 667–678, 2021b.

Wenbo Li, Guanzhong Zeng, Juncheng Zhang, Yan Xu, Yang Xing, Rui Zhou, Gang Guo, Yu Shen, Dongpu Cao, and Fei Wang. Cogemonet: A cognitive-feature-augmented driver emotion recognition model for smart cockpit. *IEEE Transactions on Computational Social Systems*, 9:667–678, 2022. URL `https://api.semanticscholar.org/CorpusID:249101490`.

Wenbo Li, Dongpu Cao, Ruichen Tan, Tianze Shi, Zhenhai Gao, Jun Ma, Gang Guo, Hongyu Hu, Jinshan Feng, and Lixin Wang. Intelligent cockpit for intelligent connected vehicles: Definition, taxonomy, technology and evaluation. *IEEE Transactions on Intelligent Vehicles*, 2023b.

Wenbo Li, Jiyong Xue, Ruichen Tan, Cong Wang, Zejian Deng, Shen Li, Gang Guo, and Dongpu Cao. Global-local-feature-fused driver speech emotion detection for intelligent cockpit in automated driving. *IEEE Transactions on Intelligent Vehicles*, 2023c.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Chenhui Liu and Wei Zhang. Learning the driver acceleration/deceleration behavior under high-speed environments from naturalistic driving data. *IEEE Intelligent Transportation Systems Magazine*, 14(3):78–91, 2020.

Hao Lu, Zitong Yu, Xuesong Niu, and Ying-Cong Chen. Neuron structure modeling for generalizable remote physiological measurement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18589–18599, 2023.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. Query rewriting for retrieval-augmented large language models. *ArXiv*, abs/2305.14283, 2023. URL `https://api.semanticscholar.org/CorpusID:258841283`.

Jiageng Mao, Yuxi Qian, Hang Zhao, and Yue Wang. Gpt-driver: Learning to drive with gpt. 2023.

Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2801–2810, 2019.

Salman Khan Muhammad Maaz, Hanoona Rasheed and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv 2306.05424*, 2023.

Xuesong Niu, Shiguang Shan, Hu Han, and Xilin Chen. Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. *IEEE Transactions on Image Processing*, 29: 2409–2423, 2019.

Geesung Oh, Euiseok Jeong, Rak Chul Kim, Ji Hyun Yang, Sungwook Hwang, Sangho Lee, and Sejoon Lim. Multimodal data collection system for driver emotion recognition based on self-reporting in real-world driving. *Sensors*, 22(12):4402, 2022.

Juan Diego Ortega, Neslihan Kose, Paola Cañas, Min-An Chao, Alexander Unnervik, Marcos Nieto, Oihana Otaegui, and Luis Salgado. Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis. In *European Conference on Computer Vision (ECCV)*, pp. 387–405, 2020.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Ambareesh Revanur, Zhihua Li, Umur A Ciftci, Lijun Yin, and László A Jeni. The first vision for vitals (v4v) challenge for non-contact video-based physiological estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2760–2767, 2021.

Mohamed H Saad, Mahmoud I Khalil, and Hazem M Abbas. End-to-end driver distraction recognition using novel low lighting support dataset. In *IEEE International Conference on Computer Engineering and Systems (ICCES)*, pp. 1–6, 2020.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. *arXiv preprint arXiv:2312.14150*, 2023.

Ronny Stricker, Steffen Müller, and Horst-Michael Gross. Non-contact video-based pulse rate measurement on a mobile service robot. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pp. 1056–1062. IEEE, 2014.

Xiaoming Tao, Dingcheng Gao, Wenqi Zhang, Tianqi Liu, Bing Du, Shanghang Zhang, and Yanjun Qin. A multimodal physiological dataset for driving behaviour analysis. *Scientific Data*, 11, 2024. URL https://api.semanticscholar.org/CorpusID:269111060.

Siyu Teng, Xuemin Hu, Peng Deng, Bai Li, Yuchen Li, Yunfeng Ai, Dongsheng Yang, Lingxi Li, Zhe Xuanyuan, Fenghua Zhu, et al. Motion planning for autonomous driving: The state of the art and future perspectives. *IEEE Transactions on Intelligent Vehicles*, 2023.

Duy Tran, Ha Manh Do, Weihua Sheng, He Bai, and Girish Chowdhary. Real-time detection of distracted driving based on deep learning. *IET Intelligent Transport Systems*, 12(10):1210–1219, 2018.

Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, et al. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*, 2023.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022.

Ziran Wang, Yougang Bian, Steven E. Shladover, Guoyuan Wu, Shengbo Eben Li, and Matthew J. Barth. A survey on cooperative longitudinal motion control of multiple connected and automated vehicles. *IEEE Intelligent Transportation Systems Magazine*, 12:4–24, 2020. URL https://api.semanticscholar.org/CorpusID:210931452.

Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. *arXiv preprint arXiv:2212.00280*, 2022.

Lin Xi, Weihai Chen, Changchen Zhao, Xingming Wu, and Jianhua Wang. Image enhancement for remote photoplethysmography in a low-light environment. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*, pp. 1–7. IEEE, 2020.

Dingkang Yang, Shuai Huang, Zhi Xu, Zhenpeng Li, Shunli Wang, Mingcheng Li, Yuzheng Wang, Yang Liu, Kun Yang, Zhaoyu Chen, et al. Aide: A vision-driven multi-view, multi-modal, multitasking dataset for assistive driving perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20459–20470, 2023a.

Jianjun Yang, Shanshan Xing, Yimeng Chen, Ruizhi Qiu, Chunrong Hua, and Dawei Dong. A comprehensive evaluation model for the intelligent automobile cockpit comfort. *Scientific Reports*, 12(1):15014, 2022.

Yi Yang, Qingwen Zhang, Ci Li, Daniel Simões Marta, Nazre Batool, and John Folkesson. Human-centric autonomous systems with llms for user command reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pp. 988–994, January 2024.

Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. Llm4drive: A survey of large language models for autonomous driving. *ArXiv*, abs/2311.01043, 2023b. URL `https://api.semanticscholar.org/CorpusID:264935408`.

Chaoyun Zhang, Rui Li, Woojin Kim, Daesub Yoon, and Paul Patras. Driver behavior recognition via interwoven deep convolutional neural nets with multi-stream inputs. *IEEE Access*, 8:191138–191151, 2020.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023a. URL `https://arxiv.org/abs/2306.02858`.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023b.

CH Zhao, BL Zhang, Jie He, and Jie Lian. Recognition of driving postures by contourlet transform and random forests. *IET Intelligent Transport Systems*, 6(2):161–168, 2012.

A Appendix

You may include other additional sections here.