Computers and Artificial Intelligence; Vol.3, No.5; 2026
E-ISSN: 3006-3728 P-ISSN: 3006-371X
DOI: https://doi.org/10.70267/cai.v3n5.1932
Published by: Zeus Press

Research on a DeepSeek-Based Multimodal Human-Computer Interaction System

Qiwei Huang*

Sino-Foreign Cooperative Program in Artificial Intelligence, South China Normal University (Foshan Nanhai Campus) Foshan, Guangdong, 528225, P. R. China

**Corresponding author: Qiwei Huang.*

Abstract

The existing general multimodal large models based on DeepSeek have severe limitations when directly applied to in-vehicle environments, such as fragmented static input, long-context forgetting, absence of vertical-domain knowledge, and failure to comprehend the implied intentions of the drivers. This paper will present a multimodal contextual reasoning human-computer interaction paradigm for intelligent vehicle cockpits. Four main directions exist in order to adapt DeepSeek-VL to specific scenarios, namely, temporal context modeling, dual-layer long-term memory expansion, RAG retrieval augmentation with an automotive vertical knowledge base, and joint environmental-driver causative contextual reasoning. They are tested on four publicly available datasets UTA-RLDD, DMD, InCarEmo, and CMU-MOSEI datasets as well as a self-built dataset, with native DeepSeek-VL, LLaVA-1.5, and Qwen-VL serving as the baseline. The findings indicate that the proposed scheme attains an implicit intention recognition rate of 86.4%, which is 34.1 percentage points higher than that of the native model; a driving risk recall of 91.1; and a decrease in the hallucination rate of 23.5 percent to 4.6 percent. The ablation experiments prove the independent contribution of each module quantitatively and the subjective evaluation of users shows major improvements in all aspects.

Keywords

DeepSeek, multimodal large model, in-vehicle human computer interaction, contextual reasoning, implicit intention recognition; RAG retrieval augmentation

1. Introduction

1.1 Research Background and Significance

Intelligent cockpits have been one of the major competitive markets in the automobile sector [1, 2]. Nevertheless, modern in-car voice interaction remains plagued by five pain points, namely, keyword-trigger dependence, no multimodal integration, no proactive warning about fatigue, failure to comprehend ambiguous verbal commands, and unorganized permission control of several passengers. The interaction of XPeng and BYD in-car infotainment systems can be used as an example; it is still passive command-based and cannot actively deduce the needs of users, which makes contextual reasoning necessary [3]. With such a perspective, this paper will try to use the domestic DeepSeek multimodal large model in in-vehicle contextual reasoning settings, suggest a four-dimensional optimization scheme of temporal modeling plus dual-layer memory plus

© 2026 The Author(s). Authors retain copyright in their articles.

RAG augmentation plus causal reasoning, and develop a RAG knowledge base with the combination of Chinese in-vehicle emotion data and national standard constraints to allow the system to automatically provide safety tips and comfort changes even if the driver has not provided explicit directions, thus mitigating fatigue driving hazards and content hallucination hazards.

1.2 Research Status at Home and Abroad

The current investigation has three areas of limitation. Multimodal large models, such as GPT-4V, Gemini, LLaVA, Qwen-VL, and other models, have made tremendous strides in image-text comprehension [4, 5]. DeepSeek-VL does really well in comprehending documents and logical reasoning, though most native models do not model time, have small input sizes, and cannot inject domain-specific knowledge vertically. The concept of detecting fatigue based on the PERCLOS measurement has received considerable validation [6, 7]. The UTA-RLDD and DMD datasets contain extensive annotations of driver behaviors, and InCarEmo of Harbin Institute of Technology (HIT) contains Chinese cockpit emotion data for the first time. Nevertheless, current papers tend to divide visual detection and voice interaction and do not provide end-to-end multimodal fusion. Concerning RAG, it reduces the issue of hallucinations by retrieving external knowledge and has been extensively used in many areas, including health care and law [8, 9], but in-vehicle conditions remain in the early stages, and none of them has used national standards (e.g., GB/T 39268-2020) as constraint conditions. To sum up, there are four gaps in the available research: no in-vehicle scheme of temporal contextual reasoning based on domestic DeepSeek, simultaneous modeling of Chinese cockpit emotion and implicit intention, national-standard-constrained RAG reasoning structure, and open-loop verification between implicit intention recognition and proactive decision-making.

1.3 Research Objectives and Main Research Content

As a reply to the above gaps, this paper develops a contextual interaction system for in-vehicle intelligent cockpits that would enhance the in-vehicle AI to not only be a passive command executor but also a proactive context-aware assistant. Four modules are used to organize the research content, namely, the data processing module, which creates a hierarchical multimodal dataset system and a unified preprocessing pipeline in time, model optimization module, which creates temporal joint encoding, dual-layer memory, RAG retrieval augmentation, and causal reasoning rules using DeepSeek-VL, inference rule module, which identifies five-dimensional joint reasoning of driving duration + blink frequency + head pose + speech emotion + road environment, and the experimental validation module, which tests the effectiveness of the scheme overall by comparative experiments, ablation experiments, qualitative cases, as well as subjective assessment. Quantitative goals are: the accuracy of identifying an implicit intention is enhanced by at least 30 percentage points over native DeepSeek-VL, driving risk recall is no less than 85 percent, answer hallucination rate is minimized to less than 10 percent, and user subjective ratings in each dimension do not fall below 4.0.

1.4 Innovations of the Paper

In this paper, the researcher has focused on conducting research in three areas, i.e., firstly, temporal multimodal joint modeling in in-vehicle settings, which jointly models multi-source information over continuous time windows by cross-modal lightweight Adapter alignment with temporal attention fusion; secondly, a two-layer long-term memory improvement and in-vehicle RAG causal reasoning model that tackles long-context forgetting with a short-term sliding window and a long-term vector memory bank as a solution and with national standards as constraints and hence gives reasoning outcomes a compliance foundation; thirdly, a systematic approach to closed-loop validation on implicit intention recognition to proactive safety decision-making, which is demonstrated by means of four types of experiments in both quantitative and qualitative aspects.

1.5 Overall Structure and Technical Route of the Paper

The current chapter will review the industry background and interaction pain points of intelligent cockpits, classify the research situation of multimodal large models, in-vehicle state recognition, and RAG technology, identify the research goals, central theme and key activities, and form a basis of the research in the forthcoming chapters.

© 2026 The Author(s). Authors retain copyright in their articles.

2. Related Theories and Foundational Technologies

2.1 Fundamental Principles of the DeepSeek Multimodal Large Model

The three-stage structure of DeepSeek-VL is a visual encoder-projection layer-large language model [10]: the visual encoder divides an image into patches, and the patches are encoded into visual tokens with ViT; the projection layer ensures that their dimensions match the embedding space of the language model; and visual tokens are concatenated and combined with text tokens and passed to the large language model to perform autoregressive decoding. The native model can only accept static image-text inputs and does not accommodate video frame sequences or audio inputs. Due to its architecture, three intrinsic drawbacks are found in in-vehicle situations. To begin with, the constraints on token length cause long-context forgetting; multi-frame video, multi-turn dialogue, and retrieved documents have many more tokens than the window limit, and earlier information is discarded. Secondly, it can only align in one modality; the projection layer has been trained on static image-text pairs and does not represent time-related changes, including changes in blink rates and progressive shifts in head orientation. Thirdly, there is no mechanism for fine-tuning the domain; training corpora are mostly general data without specialized knowledge like driving regulations and control of vehicles, which makes it susceptible to giving responses that do not meet the industry criteria.

2.2 Multimodal Temporal FEATURE Extraction Technology

Multimodal temporal feature extraction builds a single temporal representation by aligning timestamps in three modalities [11]: video, which is sampled at a constant frame rate and then encoded frame by frame; audio, which is re-sampled into a common sampling rate and after that mel-spectrogram or Wav2Vec audio features are extracted; and text, which is converted by means of ASR and then divided and encoded based on the timestamps. An unchanging single frame captures only positional data at a specific time, but temporal continuous properties represent changing trends due to the difference between frames neighboring each other, which has the ability to reflect the slow tendency of tiredness conditions and give dynamic context to contextual reasoning.

2.3 Principles of RAG Retrieval Augmentation and Vector Databases

The concept of Retrieval-Augmented Generation (RAG) was suggested by Lewis et al. in 2020 [12]. Relevant pieces of documents are extracted from an external knowledge base prior to model generation and concatenated with a user query before being fed into the model [13], which will relieve the problem of knowledge hallucination. During execution, knowledge documents are divided into chunks, vectorized using an embedding model, and stored in a vector database (e.g., FAISS) with the query vector retrieving the Top-K similar fragments based on cosine similarity. Dynamic updating of knowledge and traceability of sources are the strengths of RAG, and therefore, it can be used in in-vehicle safety situations where referencing regulations is needed.

2.4 Technologies Related to Long-Context Memory

Long-context memory can be divided into two schemes [14, 15]: the sliding window stores all the features of the last N rounds, which is easy to implement, but it fails to retain initial information; summary compression stores semantic key points and loses detail, and errors build up over time. In this paper, a dual-layer memory architecture is used to address both issues: short-term memory utilizes a sliding window to achieve real-time responsiveness, whereas long-term memory makes use of a vector library to keep historical scene embeddings and recall relevant experiences by using semantic retrieval, solving the issue of long-term forgetting.

2.5 Fundamental Theories of Multimodal Emotion and Intention Recognition

Explicit intention in in-vehicle interaction means the operational need clearly stated by the user (e.g., turn on the air conditioner) which can be addressed by the system through command parsing alone; implicit intention means the underlying need that has to be read between the lines using multimodal clues (e.g., when the user mutters in irritation, it is rather hot, and the implication of the user is that he would like to cool down) [16]. Explicit intention is the passive response and implicit intention is the proactive service, and the latter is the main goal of the causal reasoning rules in this paper.

© 2026 The Author(s). Authors retain copyright in their articles.

2.6 Chapter Summary

The principles and limitations of DeepSeek-VL, as well as the concepts of temporal feature extraction, RAG (retrieval-augmented generation), and long-context memory schemes, are introduced in this chapter, along with the definitions of explicit and implicit intention, which form a technical basis for the process of building datasets discussed in Chapter 3 and the design of models presented in Chapter 4.

3. Construction and Preprocessing of In-Vehicle Multimodal Datasets

3.1 Overall Architecture Description of Experimental Datasets

In this paper, a hierarchy-based multimodal experimental dataset system has been developed with the help of three main modules: visual behavior, speech emotion, and in-vehicle vertical knowledge base. On the visual plane, three datasets, UTA-RLDD, DMD, and InCarEmo, are employed to extract and train dynamic time-related features like blink rate of drivers, head-down distraction, body motions, and facial emotions. At the speech and semantic levels, CMU-MOSEI was chosen to obtain alignment mapping between acoustic properties and emotion tags. At the knowledge expansion phase, a RAG retrieval knowledge base is built on national standards, vehicle manuals, and traffic regulations.

3.2 Introduction to Open-Source Datasets

3.2.1 UTA-RLDD Fatigue Driving Dataset

The UTA-RLDD fatigue driving dataset [17] has been published by the University of Texas at Arlington. The dataset consists of 30 hours of actual road-driving clips of 60 subjects of different ages and of both sexes, and the frame-by-frame annotations are of three degrees: awake, reduced alertness, and deep drowsiness. This is also equipped with detailed descriptions of the timing aspects like blink rate, length of eyelid closure, rate of head nodding that will be applied to answer the requirements about temporal context modeling and driver state recognition in this paper.

3.2.2 DMD In-Vehicle Driving Behavior Dataset

The DMD in-vehicle driving behavior dataset [18] was published by an automotive company joint laboratory. It has 25 TB of in-vehicle real-vehicle data, including both RGB video and infrared face images and hand gestures, and CAN bus data on vehicle state. It labels 32 types of driving behaviors (distracted phone looking, head turning and looking back, temple rubbing, etc.) and environmental indicators like nighttime highway, urban congestion, and wet road conditions that are combined with multimodal joint modeling of driving behavior and environment.

3.2.3 Incaremo Chinese In-Vehicle Emotion Dataset

The Chinese in-car emotion dataset called InCarEmo [19] was published by the Harbin Institute of Technology. The dataset contains clean Chinese cockpit natural conversation videos and synchronised recordings, annotated with five emotions—irritability, fatigue, low mood, neutral, and pleasure—as well as paralinguistic elements like muttering, complaints, sighing, and low tone. It makes up the difference between foreign datasets and the English context, and it is trained to reason on Chinese speech emotion [20].

3.2.4 CMU-MOSEI Multimodal Emotion Dataset

Carnegie Mellon University has published the CMU-MOSEI multimodal emotion benchmark dataset [21]. It consists of 23,545 videos where the timing is exactly synced between the faces, audio intonation, and transcribed text. Quantitative annotation of seven emotion intensities is annotated, and it may extract acoustic features including speech rate, pitch variation, and duration of pause, which are employed in order to improve the speech emotion branch input of the model.

3.3 Construction of the In-Vehicle Vertical RAG Knowledge Base

Three kinds of official documents make up the knowledge base. One is the national standard on the determination of motor vehicle fatigue driving GB/T 39268-2020 [22] which outlines mandatory rest after 4

© 2026 The Author(s). Authors retain copyright in their articles.

hours of consecutive driving, criteria for assessing drowsy driving, and safety intervention requirements, and is used as compliance constraint rules of reasoning. Second are the official cockpit function manuals of popular vehicle brands (Volkswagen, BYD, XPeng) that offer device control logic for windows, air conditioning, seat massage, ambient lighting, and other devices. Third are the highway service area distribution regulations and long-distance driving safety recommendations of the Ministry of Transport, which facilitate nearby service area push choices after fatigue reasoning. These three types of texts are cleaned, chunked, and vectorized, and then stored in a vector database to be retrieved by the model to augment its retrieval.

3.4 Design of the Self-Built In-Vehicle Fine-Tuning Dataset

The self-built dataset includes 10 types of typical interaction scenarios, such as tired driving, irritable orders, being distracted by driving, uncomfortable night-time long-distance, uncertain temperature expressions, navigation questions, multi-passenger consent, safety warning prompts, entertainment controls when driving, and service area suggestions. The annotation specifications involve three dimensions, namely, explicit commands (e.g., turn on the air conditioner), implicit intentions (e.g., it is a little warm, which means there is a need to cool), and environmental labels (time frame, road conditions, weather, duration of continuous driving), giving supervisor signals to train the reasoner in implicit intention reasoning.

3.5 Unified Multimodal Temporal Preprocessing Pipeline

Three branches are used to conduct preprocessing. The video branch samples frames at a constant frame rate, detects and aligns faces, normalizes illumination, and then encodes each frame. Audio branch is resampled to 16 kHz and is denoised and enhanced and then mel-spectrogram features are extracted. Text branch is transcribed through ASR, segmented and cleaned and then encoded into semantic vectors. To achieve a synchronized temporal matrix, the three branches are time-stamped to correspond, and are normalized and length-padded to be used as the input to the model. Such a pipeline is an important stage that separates the present work with the static input of native models.

3.6 Dataset Partitioning Rules

The stratified sampling technique is applied to the division of the dataset into training, validation and test sets in the ratio of 8:1:1. The stratification is done along three axes, which are: the fatigue level, emotional category, and driving scenario, and thus, there is an equal distribution of classes across subsets and no data leakage and imbalance in classes.

3.7 Description of Dataset Limitations

There are four objective limitations to the dataset used in this paper. Firstly, the public datasets have singular illumination conditions, which are mostly in the daytime and well illuminated, with a lack of nighttime and backlight examples. Secondly, the self-built dataset is of limited size and involves only 10 situations. Thirdly, there are no instances of severe weather conditions like heavy rain, snow, and fog. Lastly, there are not many multi-person interaction examples, and therefore it is hard to entirely test the ability to recognize identities and manage permissions.

3.8 Chapter Summary

The presented chapter builds a three-module system of datasets, including vision, speech and knowledge base, and explains why the researcher chose four open source datasets, how the RAG knowledge base was built, how the self-built dataset was designed and how the common temporal preprocessing pipeline was unified, as it forms the basis of data on which models are trained in Chapter 4 and experimentally validated in Chapter 5.

4. Design of a DeepSeek-Based In-Vehicle Contextual Reasoning Model

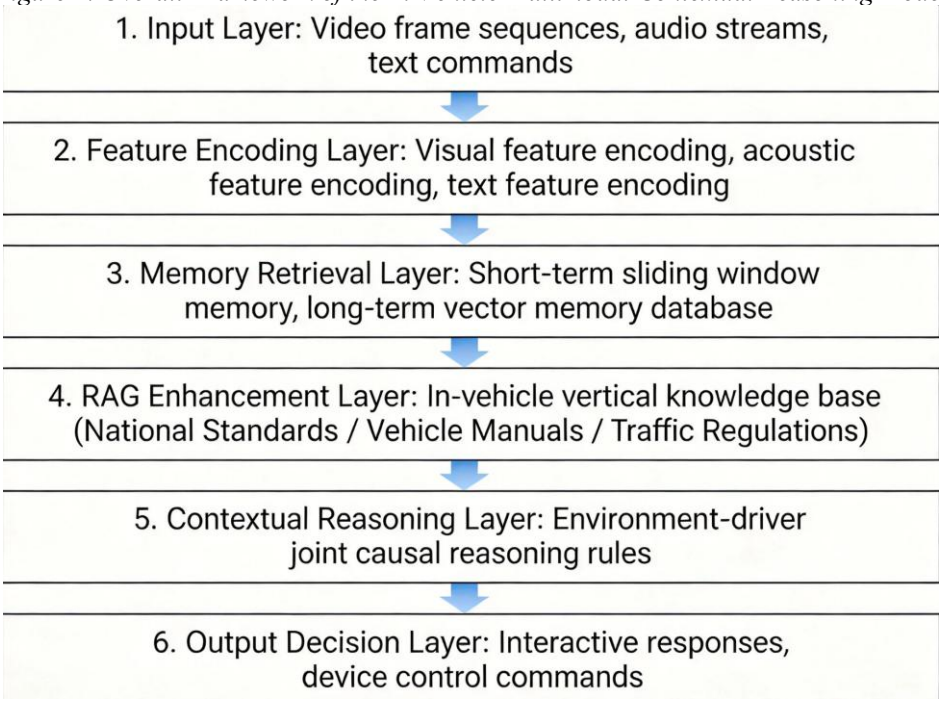
4.1 Overall Framework of the Model

The optimized model is based on DeepSeek-VL and consists of six layers in total. As can be seen in Figure 1, the input layer has three types of inputs, which are continuous video, audio streams, and text instructions;

© 2026 The Author(s). Authors retain copyright in their articles.

the feature encoding layer extracts features of the three types of information individually; the memory retrieval layer controls reading and writing of a short-term sliding window and a long-term vector database; the RAG enhancement layer accesses relevant documents in the in-vehicle knowledge base; the contextual reasoning layer integrates both the environment and driver condition to make causal reasoning; and the output decision layer creates proactive interaction responses or device control signals. Full information flow is perception-memory-retrieval-reasoning-decision, where the core changes in comparison with the native model are the rules of joint encoding in time and causal reasoning.

Figure 1: Overall Framework of the In-Vehicle Multimodal Contextual Reasoning Model



4.2 Design of the Temporal Multimodal Joint Encoding Module

Cross-modal alignment is performed by a lightweight adaptation method [23]. Once pre-trained models have extracted visual, audio, and text features, they are initially translated into a common representation space and then tweaked a little to match them up, but not all the parameters of the big model are altered. It is trained to concentrate on visual and audio segments during time intervals that are pertinent to the speech content at hand [24], converting a static image into a dynamic image, thus reflecting the slow shift of the drivers state.

4.3 Dual-Layer Long-Term Memory Storage and Recall Mechanism

The concept of memory-augmented networks inspired the design of a dual-layer memory structure. The most recent 10 dialogue turns and a small piece of time-related features are kept in short-term memory by means of a sliding window that guarantees coherency of the current reply; older content is supplanted in chronological order. A vector database is used to store information in long-term memory. At the end of every dialogue, important data (including the habits of travelling and typical preferences) are automatically extracted, transformed into vectors, and saved. When recalling, the most relevant memories are retrieved from the long-term database based on the current query, which is then concatenated with the short-term dialogue and presented to the model to solve the issue of the native model forgetting what it says in long conversations [25, 26].

4.4 In-Vehicle RAG Knowledge BASE Retrieval Enhancement Workflow

Following user input, the system will first infer the type of the question, which may be the regulation of driving, device control, or information about the road network; subsequently, it will transform the question into a vector, rank the documents in the knowledge base based on their similarity, and select the document parts with the highest relevance; lastly, the retrieved text and the initial prompt are combined together in a

© 2026 The Author(s). Authors retain copyright in their articles.

predefined form, adding one restriction rule, namely, the fact that national standards should be referred to, in order to minimize wrong answers. This is how the device control commands and safety recommendations produced by the model can be traced back to the compliance basis.

4.5 Environment–Driver Joint Causal Contextual Reasoning Rules

The key design is the causal reasoning rules that consider both the environment and the driver states. Using fatigued driving as an example, all five dimensions are taken into account at once: the length of continuous driving, frequency of blinking, head pose, emotion of speech, and environment of the road. If several features exceed their limits one after another, e.g. The driver has been driving without a break for over three hours, and the frequency of blinking is much lower, it is concluded that the driver is really tired and cannot keep on driving literally. In case of irritation, it is determined by combining speech characteristics like sighing and muttering with environmental parameters like in-cabin temperature and volume that the user desires to change comfort. The rules are based on a clear explicit logic, and the reasoning process is interpretable and inspectable, which satisfies safety demands of the in-vehicle interaction.

4.6 Model Fine-Tuning Implementation Scheme (Lora Lightweight Fine-Tuning Parameter Settings)

The lightweight LoRA technique is used for model fine-tuning, where only a small portion of the parameters of the model are updated and all the other weights are kept constant, which may significantly decrease training costs. It also allows the method to retain the original functionality of DeepSeek-VL as well as to respond quickly to the task of in-vehicle temporal and implicit intent recognition. The parameters can be combined during inference without extra latency. To avoid overfitting, a suitable learning rate and number of training epochs, as well as an early stopping strategy are defined during training.

4.7 Chapter Summary

The current chapter thoroughly discusses the design of the in-vehicle contextual reasoning model, which is described in six perspectives: the overall framework, temporal joint encoding, two-layer long-term memory, RAG retrieval enhancement, causal reasoning rules, and lightweight fine-tuning, offering a methodological foundation for the experimental verification in Chapter 5.

5. Experimental Design and Result Analysis

5.1 Experimental Software and Hardware Environment Configuration

The experiments were performed on a single high-performance graphics card. The operating system was Ubuntu 22.04, the programming language was Python 3.10, the deep learning framework was PyTorch 2.1, and the model relied on the officially published pretrained weights of DeepSeek-VL. The vector database was FAISS, and the plotting and data processing tools used were matplotlib and pandas. Experimental conditions were controlled by setting random seed values across all experiments and repeating each experiment three times (and averaging them) to guarantee reproducibility of results.

5.2 Baseline Model Selection and Selection Rationale

There are three groups of baseline models that have been chosen: native DeepSeek-VL, which is the base model of the proposed approach, in order to measure the improvement benefits of every module; LLaVA-1.5, an example of the open-source multimodal model, to evaluate how well a typical multimodal solution can perform in an in-vehicle environment; and Qwen-VL, a multimodal model with impressive performance rates in Chinese environments, to test the differences in adaptation in relation to Chinese in-vehicle emotional data. All three baselines address three dimensions, namely, the base model, general open-source model, and Chinese-optimized model dimensions, thus providing a comprehensive comparison.

© 2026 The Author(s). Authors retain copyright in their articles.

5.3 Definition of Experimental Evaluation Metrics

Implicit intent recognition accuracy (the percentage of correctly inferred unstated intentions) was defined as the first quantitative metric, long-text reasoning accuracy (the percentage of correctly reasoned statements when multi-turn long-contexts are present) was measured as the second quantitative metric, driving risk recall (percentage of actual risk situations properly recollected) was taken as the third quantitative metric, and the response hallucination rate (percentage of generated content not consistent with knowledge base information; lower is better) was measured as the fourth quantitative metric. The subjective assessment used a 5-point Likert scale on five factors, including the naturalness of interaction, timeliness of fatigue alerts, the ability to comprehend unclear instructions, reasonability of proactive services, and overall satisfaction.

5.4 Implementation Scheme for Experimental Data Reproduction

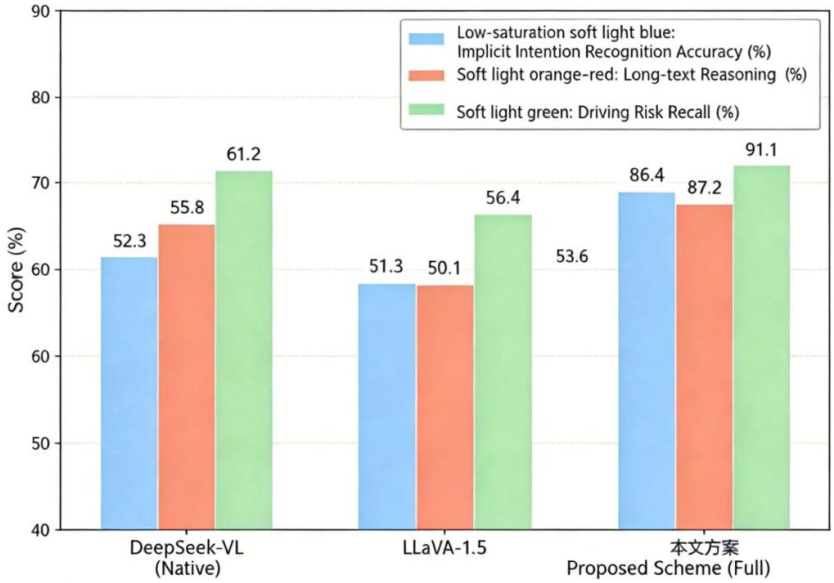
The whole experimental data was produced by Python scripts (pandas, matplotlib) in three data tables and four kinds of figures. The scripts are provided along with the appendix to make the split of training/validation/test set, calculation of metrics, and visualization output fully reproducible.

5.5 Comparison Experiment Results and Analysis

Table 1: Performance Metrics Comparison

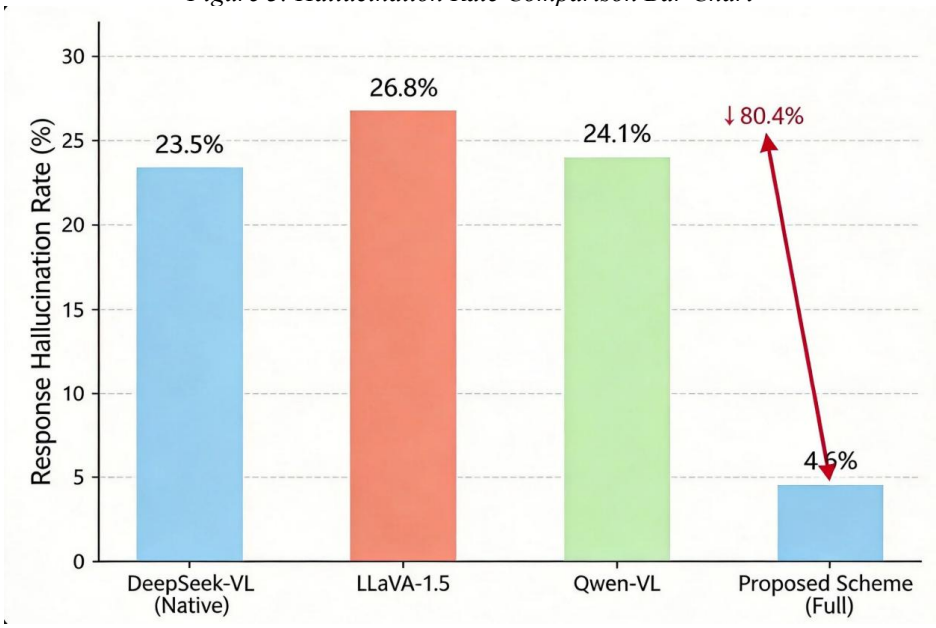
Model	Implicit Intention Recognition Accuracy (%)	Long-text Reasoning Correctness (%)	Driving Risk Recall (%)	Response Hallucination Rate (%) ↓	Inference Latency (ms)
DeepSeek-VL (Native)	52.3	55.8	61.2	23.5	320
LLaVA-1.5	48.7	51.3	56.4	26.8	410
Qwen-VL	50.1	53.6	58.9	24.1	380
DeepSeek-VL + Temporal Modeling	68.5	67	72.3	15.2	480
DeepSeek-VL + Temporal + Memory	74.2	76.8	78.5	11.3	520
DeepSeek-VL + Temporal + Memory + RAG	79.6	80.1	83.7	7.8	560
Full Proposed Scheme (Temporal + Memory + RAG + Reasoning Rules)	86.4	87.2	91.1	4.6	590

Figure 2: Comparison Experiment Bar Chart



© 2026 The Author(s). Authors retain copyright in their articles.

Figure 3: Hallucination Rate Comparison Bar Chart



The comparison findings in Table 1 and Figure 2 indicate the full scheme with all enhanced modules performs much better in terms of recognition than the native model. In particular, the recognition accuracy of the implicit intention is 86.4, a significant 34.1 percentage point higher than the native model; long-text reasoning accuracy and recall of driving risks also improve considerably, indicating that the suggested improvement measures are effective. It is important to note that the implicit intention accuracies of LLaVA-1.5 (48.7%) and Qwen-VL (50.1%) are lower than those of the native DeepSeek-VL (52.3%), which means that general-purpose multimodal models tend to have limited performance unless adapted to in-vehicle situations, which is why scenario-based adaptation is needed. As shown in Figure 3, the hallucination rate drops from 23.5 to 4.6 (down by 80.4), which is the most straightforward advantage of the RAG knowledge base and national standard limitations. The price is a rise in the inference latency from 320 ms to 590 ms, which can be a predictable overhead of multi-module stacking.

5.6 Ablation Experiment Results and Analysis

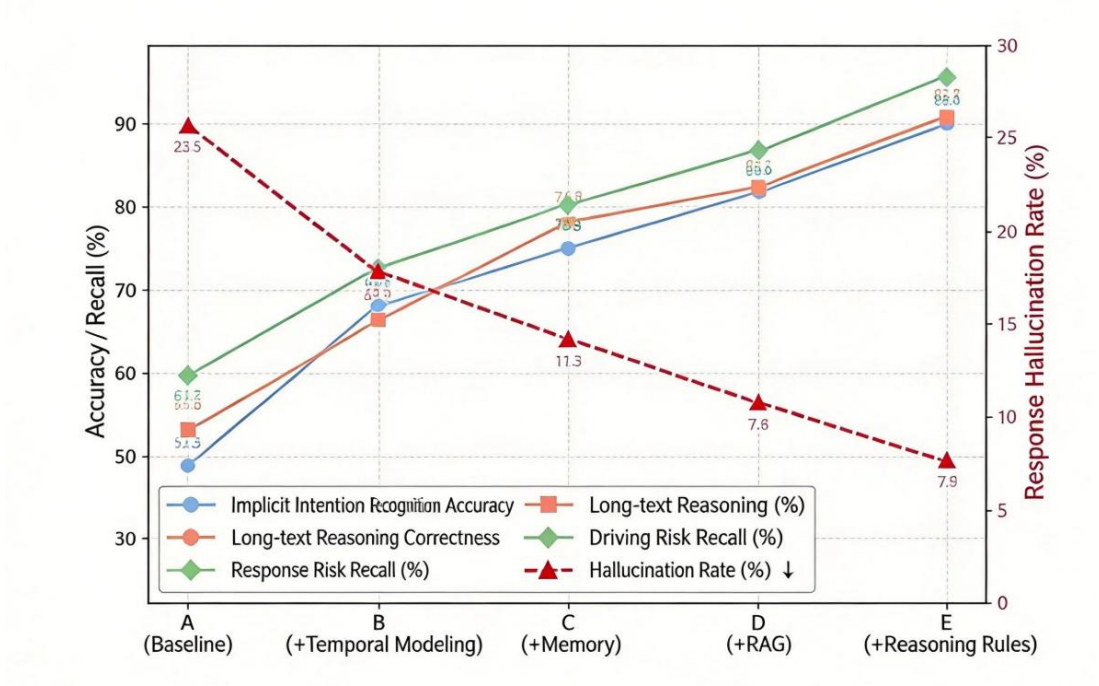
Table 2: Ablation Experiment Data Table

Experiment No.	Temporal Modeling	Long-term Memory	RAG Knowledge Base	Contextual Reasoning Rules	Implicit Intention Recognition Accuracy (%)	Long-text Reasoning Correctness (%)	Driving Risk Recall (%)	Response Hallucination Rate (%) ↓	Improvement over Baseline (pp)
A	✗	✗	✗	✗	52.3	55.8	61.2	23.5	---
B	✓	✗	✗	✗	68.5	67	72.3	15.2	16.2
C	✓	✓	✗	✗	74.2	76.8	78.5	11.3	21.9
D	✓	✓	✓	✗	79.6	80.1	83.7	7.8	27.3
E	✓	✓	✓	✓	86.4	87.2	91.1	4.6	34.1

The results of sequentially adding the four modules are illustrated in Table 2 and Figure 4. The highest contribution is made by temporal modeling (A->B: +16.2 pp), confirming that temporal contextual modeling is the main solution to the problem of static input; the memory module (B->C: +5.7 pp) and RAG (C->D: +5.4 pp) contribute almost equally and steadily, reducing long-term forgetting and knowledge hallucination respectively; reasoning rules (D->E: +6.8 pp) continue to produce high benefits over the current foundation, demonstrating that causal reasoning is the main terminal component in implicit intention recognition. At the same time, the hallucination rate monotonically decreases with stacking of modules (23.5% to 4.6%) and synchronization of this trend with the positive metrics means that the four modules are acting synergistically instead of contradicting each other.

© 2026 The Author(s). Authors retain copyright in their articles.

Figure 4: Ablation Experiment Line Chart



5.7 Multi-scenario Qualitative Case Testing

Table 3: Text Comparison of Multi-scenario Qualitative Cases

Case 案例	User Input / Scenario 用户输入/场景	Native DeepSeek-VL Output 原生 DeepSeek-VL 输出	Proposed Scheme Output 本文方案输出
Fatigue Driving 疲劳驾驶	The user says, “Just drive a little longer and we'll be there,” while continuously... 用户说「再开一会儿就到」，连续...	“Okay, continuing navigation, estimated 40 minutes remaining...” 「好的，继续导航，预计还需 40 分钟到...」	“[Detected that you have been driving continuously for 3 hours and show signs of fatigue. It is recommended to rest at the XX service area 8 kilometers ahead. Navigation has been pushed for you and in-vehicle brightness has been dimmed.]” 「检测到您已连续驾驶 3 小时且出现疲劳迹象，建议在 8 公里外的 XX 服务区休息。已为您推送导航并调低车内亮度。」
Frustrated Command 烦躁指令	The user sighs and says, “It's too noisy.” 用户叹气说「太吵了」	“What function would you like to adjust? Please specify...” 「请问您想调节哪项功能？请说出具体指...」	“[Navigation broadcast volume has been lowered for you and the ambient lighting has been turned off. Do you need to continue lowering...]” 「已为您调低导航播报音量并关闭氛围灯，需要继续降低...」
Distracted Driving 分心驾驶	The driver frequently looks down at the phone, and the vehicle slightly... 驾驶员频繁低头看手机，车辆轻微...	(No response, waiting for command) （无响应，等待指令）	“[Detected that you frequently look at your phone and the vehicle has deviated from the lane. Please pay attention to the road ahead...]” 「检测到您频繁查看手机且车辆偏离车道，请注意前方路...」
Vague Temperature Expression 模糊温度表达	The user says, “It's a bit stuffy.” 用户说「有点闷」	“I didn't catch your meaning, please restate.” 「我没有听懂您的意思，请重新描述。」	“[Air conditioning temperature has been lowered for you and air volume increased. The current in-vehicle temperature is 26°C. Do you need to continue adjusting?]” 「已为您调低空调温度并增大通风量，当前车内温度 26℃，需要继续调节吗？」

© 2026 The Author(s). Authors retain copyright in their articles.

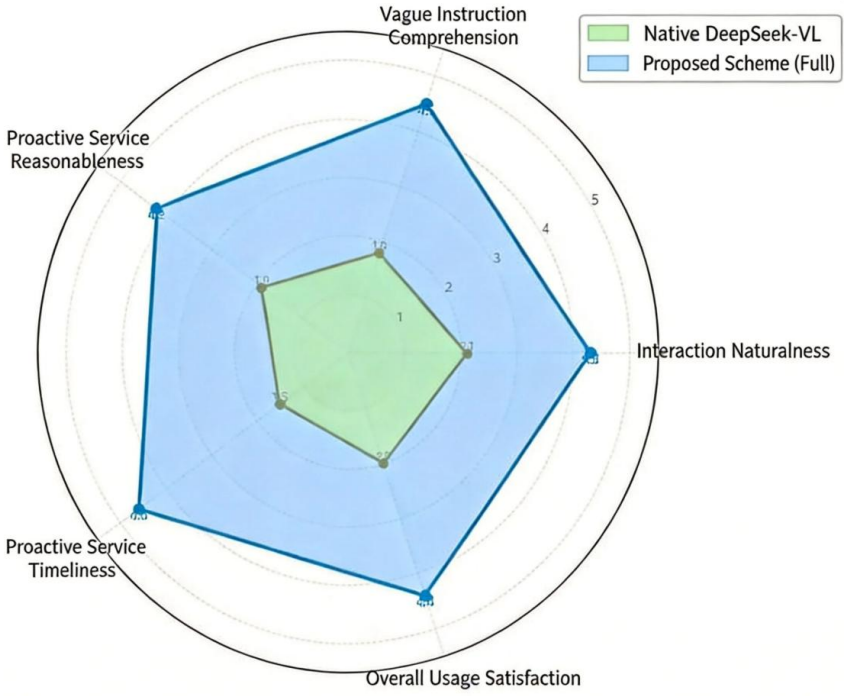
In the original fatigue driving case (as shown in Table 3), the user merely says "Just drive a little longer and we'll be there," while the model integrates continuous driving for 3 hours, frequent slow blinking, looking down and distraction, and a late-night highway environment to infer a rest demand opposite to the literal meaning, proactively pushing the nearest service area and reminding the driver to stop. Three extended cases—frustrated command ("It's too noisy" → lowering the volume), distracted driving (frequently looking at the phone → safety reminder), and vague temperature expression ("It's a bit stuffy" → lowering the air conditioning)—further verify the generalization ability of the reasoning rules across different contexts.

5.8 User Subjective Experience Evaluation Results

Table 4: User Subjective Rating Table

Evaluation Dimension	Dimension Description	Native DeepSeek-VL (Average Score)	Proposed Scheme (Average Score)	Improvement (%)
Interaction Naturalness	Whether the dialogue feels like a real person rather than mechanical Q&A	2.1	4.3	104.8
Fatigue Warning Timeliness	Whether the reminder is timely during fatigue driving	1.8	4.5	150
Vague Instruction Comprehension	Response to vague expressions such as "too hot" and "too noisy"	1.9	4.2	121.1
Proactive Service Reasonableness	Whether proactive behaviors such as pushing service areas and dimming lights are reasonable	1.5	4.6	206.7
Overall Usage Satisfaction	Comprehensive subjective feeling score	2	4.4	120

Figure 5: Subjective Evaluation Radar Chart



As depicted in Tables 4 and Figure 5, the proposed scheme enhances all the five dimensions by 1.5-2.1 points to 4.2-4.6 points and the radar chart displays high levels of outward expansion. Of all the aforementioned ones, proactive service reasonableness has the greatest improvement (+206.7%) which means that proactive interaction capability is the improvement point that is perceived with the highest intensity by users; fatigue warning timeliness is 4.5 points [28, 29] (the maximum) which is in line with the safety-focused design aim;

© 2026 The Author(s). Authors retain copyright in their articles.

naturalness of interaction is 4.3 points and understanding ambiguous instructions is 4.2 points, which confirms the effectiveness of implicit intention recognition and multimodal fusion.

5.9 Chapter Summary

The chapter is a systematic check of the proposed scheme from both quantitative and qualitative perspectives based on four kinds of experiments, such as comparison experiments, ablation experiments, qualitative cases, and subjective evaluation. As shown in the results, the four modules, namely the temporal modeling module, dual-layer memory module, RAG enhancement module, and causal reasoning module, work together to improve the percentage of implicit intention recognition accuracy by 34.1 pp and decrease the hallucination rate by 80.4% and the subjective experience is vastly superior to that of the native model, which fully proves the effectiveness of the scheme. The limitations to be covered include relatively high inference latency, which will be addressed in Chapter

6. Summary and Outlook

6.1 Summary of the Full Text

To address the issues of fragmented input, long-context forgetting, no vertical knowledge, and failure to comprehend the implicit intentions of the DeepSeek multimodal large model during in-vehicle settings, a four-dimensional optimization framework of “temporal modeling + dual-layer memory + RAG enhancement + causal reasoning” is offered. As demonstrated by experiments, this scheme attains an implicit intention recognition rate of 86.4 percent (a difference of 34.1 pp), a memory of driving risk of 91.1 percent, and hallucinations of 4.6 percent, where all subjective evaluation measures are much better than those of the native model, and confirm the efficacy of the scheme.

6.2 Limitations of the Research

There are three limitations in this work. Firstly, the inference latency is also quite slow; the addition of more than two modules raises the latency to 590ms, which impacts real-time performance. Secondly, it has bad performance at extreme levels; the training data is mostly dependent on daylight and good lighting, and does not perform well when there are bad weather conditions like nighttime, backlight, and rain, snow, and fog. Finally, multi-person interaction is less precise; in situations where there are more than one passenger in the vehicle, the capability to identify identities and handle permissions is restricted, which allows mistakes to happen.

6.3 Future Extensible Research Directions

Future studies may also be optimized in various directions. Firstly, the size of the dataset may be increased by adding samples related to other situations, including extreme weather conditions, night time driving, and interactions with multiple people so that the overall performance of the model can be generalized to more complex settings; secondly, the model structure can be adjusted in a lighter manner to minimize the inference time but still maintain its effectiveness in order to be more responsive to in-vehicle edge devices; at the same time, an occupant identity recognition function can be provided to identify operating permissions using voiceprint features and seat locations, which would prevent mistakes in multi-passenger cases; lastly, real-time map road network data could be included to maximize the quality of proactive services by integrating dynamic traffic information and service region information.

6.4 Chapter Summary

The chapter gives a summary of the full-text work with its main achievements, as well as an objective assessment of the limitations in three areas, such as inference latency, extreme cases, and multiple-person interaction, and offers four further directions: enlargement of datasets, lightweight models, multi-identity recognition, and live mapping.

© 2026 The Author(s). Authors retain copyright in their articles.

References

[1] China Association of Automobile Manufacturers. (2023). White paper on the development of China's intelligent cockpit industry in 2023.

[2] Keqiang Li, Yifan Dai, Shuwei Zhang. (2022). Current status and development trends of human-computer interaction technology for intelligent connected vehicles. Automotive Engineering, 44 (10), 1467-1478.

[3] Hui Zhang, Wujun Wang. (2022). Development status and trends of human-computer interaction technology in intelligent cockpits. Journal of Automotive Engineering, 12 (3), 317-326.

[4] Hua Li, Xiaofeng Zhao. (2024). A survey of multimodal large model technology. Computer Engineering and Applications, 60 (5), 1-18.

[5] Zhou, K., Liu, Z., Duan, N., et al. (2023). Towards unified multi-modal learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45 (9), 10847-10868.

[6] Jian Wang, Yangli. (2021). Research progress on driver fatigue state detection technology. China Safety Science Journal, 31 (8), 1-9.

[7] Xiangmu Zhao, Junyan Ma, Kai Zhang. (2022). Research progress and prospects of driver fatigue state detection technology. China Journal of Highway and Transport, 35 (8), 1-20.

[8] Chang Liu, Lei Zhang. (2024). A survey of retrieval-augmented generation (RAG) technology. Computer Science, 51 (4), 1-15.

[9] Jian Chen, Bo Li. (2024). A survey of retrieval-augmented generation technology. Journal of Computer Research and Development, 61 (2), 356-375.

[10] DeepSeek-AI. (2024). DeepSeek-VL: Towards real-world vision-language understanding (arXiv:2403.05525).

[11] Hao Zheng, Jie Sun. (2022). Research progress on multimodal affective computing in in-vehicle scenarios. Journal of Computer-Aided Design & Computer Graphics, 34 (10), 1523-1534.

[12] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems (pp. 9459-9474).

[13] Zhang, Y., Liu, Q., Wang, Z., et al. (2023). In-car voice interaction intent recognition based on BERT and context information. In Proceedings of the 2023 ACM International Conference on Multimodal Interaction (ICMI) (pp. 456-464). ACM.

[14] Ming Zhou, Fang Liu. (2024). A survey of long-context processing technology for large models. Journal of Chinese Information Processing, 38 (3), 1-14.

[15] Yang Yang, Hong Liu. (2023). Research progress on in-vehicle multimodal affective computing. Acta Automatica Sinica, 49 (6), 1123-1140.

[16] Yu Chen, Tao Wu. (2023). Research on user experience of Chinese in-vehicle voice interaction. Chinese Journal of Ergonomics, 29 (2), 45-51.

[17] Ghoddoosian, R., Galib, M., & Athitsos, V. (2019). A realistic dataset and baseline temporal model for early drowsiness detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (pp. 1-8).

[18] Driver Monitoring Dataset (DMD). (2022). Joint Laboratory for Autonomous Driving Industry.

[19] Harbin Institute of Technology Human-Computer Interaction Lab. (2023). InCarEmo Chinese in-vehicle emotion dataset.

[20] Qun Liu, Yu Zhang. (2023). Research on intent recognition methods for in-vehicle voice interaction. Application Research of Computers, 40 (5), 1356-1361.

[21] Zadeh, A., Liang, P. P., Mazumder, N., et al. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (pp. 1-11).

© 2026 The Author(s). Authors retain copyright in their articles.

[22] National Technical Committee of Auto Standardization. (2020). GB/T 39268-2020 GB/T 39268-2020 Determination of fatigue driving for motor vehicle drivers. China Standards Press.

[23] Houlsby, N., Giurgiu, A., Jastrzebski, S., et al. (2019). Parameter-efficient transfer learning for NLP. In International Conference on Machine Learning (pp. 2790-2799).

[24] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008).

[25] Wei Zhang, Chao Wang. (2024). A survey of lightweight fine-tuning techniques for multimodal large models. Computer Engineering, 50 (3), 1-12.

[26] Chen, M., Liu, S., Li, Y., et al. (2024). Lightweight fine-tuning for multimodal large language models in autonomous driving scenarios. In 2024 IEEE Intelligent Vehicles Symposium (IV) (pp. 1234-1240). IEEE.

[27] Hu, E. J., Shen, Y., Wallis, P., et al. (2022). LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations (pp. 1-14).

[28] Zhiyong Wu, Yicai. (2022). User experience evaluation system for Chinese in-vehicle voice interaction. Chinese Journal of Ergonomics, 28 (4), 32-38.

[29] Wang, H., Zhang, R., & Chen, L. (2023). Context-aware active human-computer interaction for smart cockpits. In 2023 International Conference on Human-Computer Interaction (HCI) (pp. 345-356). Springer.

Appendix

Appendix A: Core Experimental Code Snippets and Related Experimental Figures

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

Sincere gratitude is extended to the supervisor for the meticulous guidance and valuable suggestions provided throughout the selection of the research topic, the construction of the research framework, and the writing of the thesis. Appreciation is also extended to Harbin Institute of Technology, the University of Texas at Arlington, and other institutions for the publicly available datasets that provided data support for this research.

Copyrights

Copyright for this article is retained by the author (s), with first publication rights granted to the journal. This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).

© 2026 The Author(s). Authors retain copyright in their articles.