VLLM Guided Human-Like Guidance Navigation Generation

Masaki Nambata[1][a], Tsubasa Hirakawa[1][b], Takayoshi Yamashita[1][c], Hirobobu Fujiyoshi[1][d],
Takehito Teraguchi[2][e], Shota Okubo[2] and Takuya Nanri[2][f]

[1]*Chubu University, 1200 Matsumoto-cho Kasugai, Aichi, Japan*

[2]*Nissan Motor Co., Ltd., 2 Takara-cho Kanawgawa-ku Yokohama-shi, Kanagawa, Japan*

{*masaknanbt, hirakawa*}*@mprg.cs.chubu.ac.jp*, {*takayoshi, fujiyoshi*}*@isc.chubu.ac.jp*, {*shota-ohkubo, t-nanri,*

Keywords: Driver's Assistance System, Vision and Language Model, Evaluation Method.

Abstract: In the field of Advanced Driver Assistance Systems (ADAS), car navigation systems have become an essential part of modern driving. However, the guidance provided by existing car navigation systems is often difficult to understand, making it difficult for drivers to understand solely through voice instructions. This challenge has led to growing interest in Human-like Guidance (HLG), a task focused on delivering intuitive navigation instructions that mimic the way a passenger would guide a driver. Despite this, previous studies have used rule-based systems to generate HLG datasets, which have resulted in inflexible and low-quality due to limited textual representation. In contrast, high-quality datasets are crucial for improving model performance. In this study, we propose a method to automatically generate high-quality navigation sentences from image data using a Large Language Model with a novel prompting approach. Additionally, we introduce a Mixture of Experts (MoE) framework for data cleaning to filter out unreliable data. The resulting dataset is both expressive and consistent. Furthermore, our proposed MoE evaluation framework makes it possible to perform appropriate evaluation from multiple perspectives, even for complex tasks such as HLG.

1 INTRODUCTION

Advanced Driver Assistance Systems (ADAS) aim to develop technologies that enhance driver safety and comfort. Car navigation systems, a key component of ADAS, have become indispensable in daily life. These systems typically generate navigation instructions based on GPS and map data, producing directions such as "Turn left at the intersection 100 meters ahead." However, distance-based guidance can be challenging to interpret using voice instructions alone, often requiring drivers to check the in-vehicle display, which leads to distractions. Recently, navigation systems have started to use Geographic Information System (GIS) data to enhance instructions by incorporating landmarks like traffic lights (e.g., "Turn left at the next traffic light"). While helpful, GIS data can quickly become outdated, causing confusion for drivers and failing to fully address the challenges of providing intuitive guidance. Outdated or inaccurate data can increase driver stress, potentially leading to accidents and traffic congestion(Barrow, 1991).

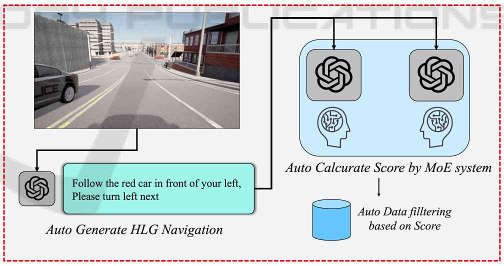

Follow the red car in front of your left,
Please turn left next
Auto Generate HLG Navigation
Auto Calcurate Score by MoE system
Auto Data fillteing
based on Score

Figure 1: Overview of the flow of the proposed method.The proposed method generates a dataset of guidance sentences using the Human-like Thought Few-shot Chain-of-Thought prompting (HLTFC) method, followed by filtering through the Mixture of Experts (MoE)-based data cleaning framework.

In contrast, the next generation of car navigation systems, HLG (Human-like Guidance), is expected to become a reality. HLG is a task aimed at providing navigation that can be intuitively understood by the driver, similar to guidance provided by a passenger. By presenting the target intersection in a way

[a] https://orcid.org/0009-0006-4903-203X
[b] https://orcid.org/0000-0003-3851-5221
[c] https://orcid.org/0000-0003-2631-9856
[d] https://orcid.org/0000-0001-7391-4725
[e] https://orcid.org/0009-0000-4719-5739
[f] https://orcid.org/0009-0001-4592-590X

Nambata, M., Hirakawa, T., Yamashita, T., Fujiyoshi, H., Teraguchi, T., Okubo, S. and Nanri, T.
VLLM Guided Human-Like Guidance Navigation Generation.
DOI: 10.5220/0013191100003912

In *Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2025) - Volume 2: VISAPP*, pages 456-463
ISBN: 978-989-758-728-3; ISSN: 2184-4321

that the driver can intuitively comprehend based on the scene in front of them, HLG addresses the issues present in existing car navigation systems. Previous research proposed an HLG dataset that included forward-facing video from the driver's perspective in a simulator, gaze information of the driver, and navigation sentences(Nambata et al., 2023). This dataset, as far as we know, is the only one used in prior HLG research. The navigation sentences in (Nambata et al., 2023) consist of expressions like, ”Turn left at the intersection where you see the red car,” where objects are used to describe the location of the intersection. It was claimed that intuitive guidance could be constructed by determining the reference object for navigation based on the driver's gaze information. However, the navigation sentences were created by defining five template sentences in advance and selecting from these templates based on contextual data collected from the simulator. Therefore, this rule-based dataset of navigation sentences lacks diversity in sentence expressions and is not easily extendable to real-world data, resulting in poor dataset quality. On the other hand, creating datasets manually involves high annotation costs. Additionally, when using large numbers of annotators, such as crowd workers, it is difficult to maintain consistency and quality in the data. To develop more efficient and accurate models, high-quality datasets with broad expressive power and minimal noise are essential(Zha et al., 2023).

To address these issues, this study proposes a method to automatically generate high-quality navigation sentence datasets solely from image data. First, we introduce a novel prompting technique that leverages GPT-4, a vision-language model (VLLM) with advanced image recognition and instruction-following capabilities, to generate scene-appropriate navigation sentences. Our method incorporates the decision-making processes of drivers receiving navigation, as analyzed by Passini et al. (Passini, 1984), and the spatial understanding processes studied by Evans et al. (Evans et al., 1984), structuring the generation steps in a Chain-of-Thought (CoT) format (Wei et al., 2022). Furthermore, we refine the few-shot learning prompts into conditional prompts (Brown et al., 2020) to ensure consistency in the generated sentences(Brown et al., 2020). By using these constructed prompts, consistent navigation sentences are automatically generated from images. However, due to the hallucination problem of LLMs, a portion of low-quality data is still generated by our prompting method. To address this issue, we propose an automatic evaluation and data-cleaning method using a Mixture of Experts (MoE) framework. This method evaluates navigation sentences from multiple perspectives by utilizing several LLMs with different reasoning processes. Multi-perspective evaluation by multiple LLMs, unreliable data is filtered out and high-quality data is automatically generated.

Our experiments confirm the high quality of the datasets generated using the proposed prompting and evaluation methods. Furthermore, we quantitatively demonstrate the impact of the proposed prompting elements on the generation outcomes. In addition, the MoE-style automatic evaluation framework can provide appropriate evaluation for complex tasks such as HLG through appropriate prompt design and multi-perspective evaluation.

- We propose a novel prompting method that mimics human thought processes, generating consistent and intuitive guidance sentences suitable for HLG.
- We introduce an automatic evaluation method for Vision & Language data using VLLMs, allowing accurate assessment of complex tasks like HLG.
- We propose a framework for the automatic generation of high quality datasets using our proposed method. It is possible to create high quality data with little effort from image data alone. In addition, it is believed that it will be possible to respond to other tasks by constructing prompts according to our method.
- We provide a dataset for the realisation of HLG generated by our method.

2 RELATED WORK

2.1 Datasets Creation Methods by LLM

To efficiently build high-performance AI models, high-quality datasets are essential. Before the advent of Large Language Models (LLMs), manual annotation was the most common method, but manual data labeling is costly and time-consuming. Annotation by crowdworkers is problematic because of individual differences in quality, making it difficult to maintain consistency(Chmielewski and Kucker, 2019). Since the emergence of LLMs, especially language models like GPT (Radford et al., 2018), which demonstrate high performance across various domains, automated annotation using LLMs has gained attention. It has already been reported that for text-to-text tasks such as translation, providing GPT with appropriate prompts can generate data of higher quality than human-annotated datasets (He et al., 2024), (Oh et al., 2023), (Yu et al., 2023). Recently, with the high multimodal reasoning capabilities of GPT-4 (Achiam et al., 2023), high-quality datasets have been created even for vision and language tasks by designing appropriate prompts (Liu et al., 2023a), (Wang et al., 2023). Liu et al. constructed a dataset using GPT-4 to build a large-scale Vision & Language model (Liu et al., 2023a). However, this dataset is designed using prompts that consist of detailed captions describing the images and bounding box information of objects within the images, representing the images solely in text. Therefore, it has limitations in its ability to capture the interaction between images and text. Wang et al. created a dataset using GPT-4 Omni (Wang et al., 2023). In addition to image data, their dataset is constructed by providing location information and object category data in a conversational format within the input prompt.

In the HLG task addressed in this study, a dataset containing driver-perspective images, driver gaze information, and navigation sentences was proposed in previous research (Nambata et al., 2023). The navigation sentences are created by selecting from five predefined template sentences, based on context data collected in a simulator. However, such rule-based navigation sentence datasets have poor textual variety and lack extensibility to real-world data, resulting in low dataset quality. In this study, following prior research by Wang et al. and other studies on automatic dataset creation, we automatically generate the dataset by designing appropriate custom prompts.

2.2 Evaluation Method for Vision & Language Tasks

Various evaluation metrics for assessing generated text in Vision & Language tasks have been extensively studied. Early metric-based methods such as BLEU, ROUGE, and METEOR, commonly used in the initial stages, calculate the degree of matching with reference data based on n-grams (Papineni et al., 2002), (Lin, 2004), (Banerjee and Lavie, 2007). However, these methods fail to capture semantic similarity and are thus unable to provide valid evaluations. After the introduction of Transformer models (Vaswani et al., 2017), embedding-based methods like BERTScore (Zhang et al., 2020) and MoverScore (Zhao et al., 2019) were developed. While embedding-based methods can capture semantic similarity, they rely on ground truth data, making it difficult to obtain highly reliable evaluations. In recent years, evaluation methods using LLMs, which do not require ground truth data, have gained attention, and methods such as GPTScore and G-EVAL have been proposed (Fu et al., 2024), (Liu et al., 2023b). However, these methods evaluate the generated text itself and are not suitable for tasks like VQA, where the relationship with image information is crucial, or for the HLG task that we are working on.

Hessei et al. proposed CLIPScore, which evaluates the similarity between images and text using CLIP(Radford et al., 2021) trained on large datasets, and confirmed that it has a high correlation with human evaluations (Hessel et al., 2021). Pranava et al. proposed VIFIDE, which calculates the similarity between object instances in images and words in text, and confirmed its high correlation with human judgment (Madhyastha et al., 2019). Both evaluation metrics can assess the semantic similarity between image information and text, but it is challenging to accurately evaluate specialized texts such as HLG.

3 PROPOSED METHOD

Previous studies on Human-Like Guidance (HLG) created datasets by generating various scenes using CARLA and producing corresponding navigation sentences. However, these sentences were based solely on context data from CARLA, which limited sentence diversity and scalability, resulting in lower dataset quality. Manual annotation, on the other hand, introduces challenges related to cost, consistency, and quality. To address these issues, this study proposes a framework to automatically generate high-quality datasets.

3.1 Proposed Method Overview

Figure 2 shows the overview of the proposed method. Our method consists of the following two steps.

As the first step, we initially create the guidance text data from images and prompts using GPT-4o. Simple prompts alone are not enough to generate appropriate guidance text. To address this issue, we propose a Human-like Thought Few-shot Chain-of-Thought prompt (HLTFC), which take in to account the characteristics of human-like navigation provided by passengers.

As the second step, we further evaluate and filter the generated data. Even if we use the proposed prompt at the previous stage, it remains difficult to create a high-quality dataset due to issues such as hallucinations inherent in LLMs. Therefore, we propose the GPT guided Auto Cleaning Framework (G-ACF). The G-ACF is based on a Mixture of Experts (MoE) framework and utilizes multiple LLMs with different input prompts.

Simple prompts, however, are insufficient for generating accurate navigation sentences. To address

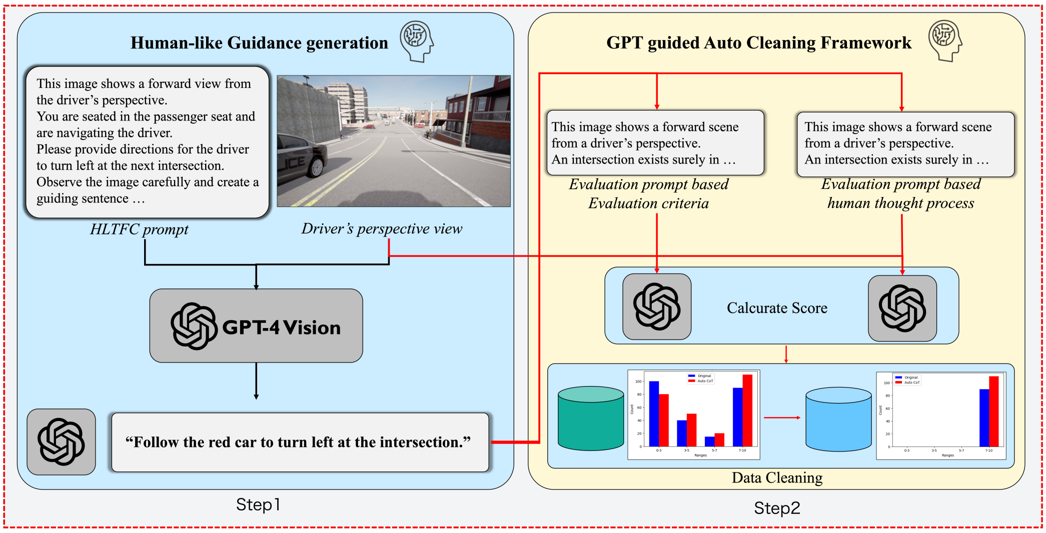

Human-like Guidance generation
This image shows a forward view from the driver's perspective.
You are seated in the passenger seat and are navigating the driver.
Please provide directions for the driver to turn left at the next intersection.
Observe the image carefully and create a guiding sentence …
HLTFC prompt
Driver's perspective view
GPT-4 Vision
"Follow the red car to turn left at the intersection."
Step1
GPT guided Auto Cleaning Framework
This image shows a forward scene from a driver's perspective.
An intersection exists surely in …
Evaluation prompt based Evaluation criteria
This image shows a forward scene from a driver's perspective.
An intersection exists surely in …
Evaluation prompt based human thought process
Calcurate Score
Data Cleaning
Step2

Figure 2: In Step 1, we generate a consistent dataset of guidance sentences using GPT-4o and the proposed Human-like Thought Few-shot Chain-of-Thought Prompt (HLTFC). In Step 2, the generated dataset is filtered using the GPT-guided Auto Cleaning Framework (G-ACF). These two steps enable the automatic creation of high-quality Vision & Language datasets.

this, we propose a Human-like Thought Few-shot Chain-of-Thought prompt (HLTFC), which takes into account the characteristics of human-like navigation provided by drivers or passengers. Despite using HLTFC, the quality of the generated dataset can still suffer from issues such as hallucinations, a common problem in large language models (LLMs). Therefore, we propose the GPT guided Auto Cleaning Framework (G-ACF), a data evaluation framework in the form of a Mixture of Experts (MoE). By employing multiple LLMs with different input prompts, this framework evaluates the generated navigation sentences from various perspectives, filtering out unreliable data to ensure high-quality outputs.

3.2 Human-Like Thought Few-Shot Chain-of-Thought Prompt

HLG aims to provide navigation sentences that drivers can intuitively understand, similar to directions given by a passenger. On the other hand, when humans grasp space, they use surrounding objects to judge distances and recognize the environment (Evans et al., 1984). Therefore, it has been proven that representing the location of target intersections using objects makes navigation sentences easier for drivers to understand (Burnett, 2000), (Allen, 1999), (Tom and Denis, 2003).

From the above, the following three points are important in constructing the HLG Navigation sentence.

- Representation of intersection locations using objects
- Appropriate objects to be used in that case
- Methods of expressing sentences

In the dataset proposed by previous research on HLG, objects representing the position of intersections were selected using driver's gaze information. However, the navigation sentences were created using five predefined template sentences, which imposed limitations on the expression of the intersection position and the navigation sentence itself. Additionally, since the template sentences were selected based on context data collected from the simulator, there is a lack of scalability to real-world data. Due to these issues, the dataset from previous research is insufficient for realizing HLG.

In response, this research creates a navigation sentence dataset that considers three important elements in HLG. To generate appropriate navigation sentences, we propose a Human-like Thought Few-shot Chain-of-Thought Prompt (HLTFC), which combines Chain-of-Thought (CoT) and Few-shot methods with improvements, and use GPT-4o, which has high image recognition and instruction-following capabilities, to create the dataset. HLTFC is constructed from a role assignments to GPT-4 Omni (GPT-4o), conditions for generating navigation sentences, and a step-by-step thought process. Through this, the LLM follows the same thought process as humans when giving directions and generates appropriate navigation sentences. First, we will give GPT-4o the role of 'sitting in the passenger seat and navigating the driver'. Next, as conditions, we instruct it to use objects to represent the position of intersections, specify the objects to be used for this representation, assume car navigation, and make the sentences concise. The objects used for this representation are, as in previous research, the objects that the driver is gazing at, based on driver gaze information. At this point, several examples of intersection representation are provided in a few-shot format. This allows for both explicit intersection expressions such as ”Turn left at the intersection where the red car is” and implicit expressions like ”Follow the red car and turn left.” Finally, as a thought process for creating navigation sentences, we present a step-by-step thought prompt based on the decision-making process of the driver receiving the directions and the human process of grasping the space. With HLTFC, the LLM can follow the same thought process as humans when giving directions, generate human-like navigation sentences, and automatically create navigation sentence data suitable for HLG.

3.3 Filtering Data by Multiple LLMs

The quality and reliability of the dataset are critical factors that influence model performance more than the model structure itself. Despite the effectiveness of HLTFC, ensuring perfect dataset quality remains challenging due to factors like hallucinations. Therefore, it is necessary to examine the quality of the dataset and clean it appropriately. However, existing data cleaning methods for datasets in Vision & Language tasks are not well-suited for the specific domain of scene context and Navigation sentences, as in this research (Xu et al., 2023), (vdc, 2024). In addition, it is difficult to appropriately evaluate the quality of the special format of the guidance text using existing evaluation metrics for Vision & Language tasks such as CLIPScore and VIFIDEL.

To address this issue, we introduce GPT-4o’s powerful multimodal reasoning capabilities to automatically evaluate and clean the dataset. However, there remain concerns about whether the automatic evaluation by LLMs is truly adequate. Therefore, in this research, we propose a GPT-guided auto cleaning framework (G-ACF), a Mixture of Experts (MoE) evaluation framework that uses multiple GPT-4os with different evaluation processes to filter the data. By employing multiple LLMs with different evaluation processes, it becomes possible to evaluate the data from multiple perspectives, G-ACF filters out unreliable labels and constructs a more diverse, high-quality dataset.

The overview of G-ACF is shown in Figure 3. In this research, we experimentally prepare two types of evaluation LLMs. Before constructiong two LLMs, based on the three points are important in constructing the HLG, we establish six evaluation criteria for assessing sentences.

- Whether the mentioned object is present
- Whether the driver’s gaze matches the object
- Whether the expression of the intersection using the object is appropriate
- Whether the mentioned direction of travel is correct
- Whether there are any expressions that could cause confusion regarding the direction of travel
- Whether the sentence length is appropriate for car navigation

For the first evaluation LLM, we followed the evaluation metrics proposed by Liu et al. and input the above defined evaluation criteria into LLM to automatically create an evaluation step prompt(Liu et al., 2023b). Through this, the LLM evaluates the navigation sentence according to the criteria.

For the second LLM, based on the decision-making process of drivers receiving navigation analyzed by Passini et al. (Passini, 1984), and the spatial grasping process of humans analyzed by Evans et al. (Evans et al., 1984), evaluation steps are constructed and input into the LLM as a prompt. Through this, the LLM evaluates the navigation sentence using a thought process similar to that of a human.

Each of these evaluation LLMs follows its respective evaluation steps and assigns a score on a scale of 0 to 10 (with one decimal point). The final score is obtained by averaging the scores from the two evaluation LLMs.

4 EXPERIENCE

We conduct experiments to confirm the effectiveness of our proposed method. First, we evaluate the quality of the dataset created by our proposed Human-like Thought Few-shot Chain-of-Thought prompt (HLTFC) and compare it quantitatively with datasets created by other prompting methods. Next, we fine-tune the VLM using the dataset created by our data cleaning method and investigate its effectiveness based on the model accuracy and output sentences.

4.1 Effectiveness Studies of Prompting Methods

We investigate the effectiveness of the proposed Human-like Thought Few-shot Chain-of-Thought Prompt (HLTFC). For comparison, we use existing standard prompting methods: Few-shot prompts and

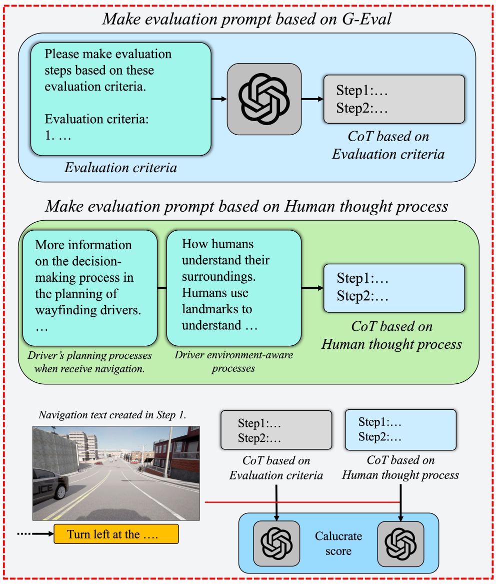

Make evaluation prompt based on G-Eval
Please make evaluation steps based on these evaluation criteria.
Evaluation criteria:
1. …
Evaluation criteria
Step1:…
Step2:…
CoT based on Evaluation criteria
Make evaluation prompt based on Human thought process
More information on the decision-making process in the planning of wayfinding drivers.
…
Driver's planning processes when receive navigation.
How humans understand their surroundings. Humans use landmarks to understand …
Driver environment-aware processes
Step1:…
Step2:…
CoT based on Human thought process
Navigation text created in Step 1.
Turn left at the ….
Step1:…
Step2:…
CoT based on Evaluation criteria
Step1:…
Step2:…
CoT based on Human thought process
Calucrate score

Figure 3: Overview of G-ACF. We evaluate the guidance text using two evaluation machines: one that evaluates according to the evaluation criteria we have constructed, and another that evaluates according to the thought processes used by the driver for planning and spatial awareness. We construct a high-quality data set by performing a multifaceted evaluation using the two evaluation machines and filtering out data with low scores.

Table 1: Quantitative comparison using each evaluation model of G-ACF. The values are the average of all data.

	Evaluation Critica prompt	Human thought prompt	Correlation coefficient
Few-shot	8.74	8.50	0.68
CoT	9.28	9.23	0.60
HLTFC (ours)	**9.58**	**9.27**	0.68

Table 2: Some of the probability of occurrence of words indicating intersections.

	Appearance probability (%)		
	Few-shot	CoT	HLTFC
Follow	35.1	0.0	15.3
with	21.4	1.5	3.2
near	7.2	6.5	12.1
past	11.5	0.9	7.
after	3.2	0.1	2.1

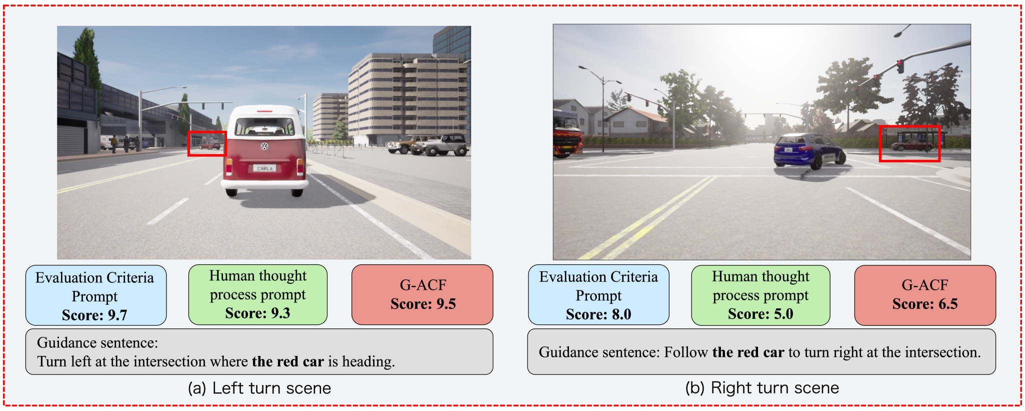
Chain-of-Thought (CoT) prompts. Each prompt is constructed by excluding elements from HLTFC. The Few-shot prompt is created by removing the CoT element from HLTFC, while the CoT prompt is constructed by excluding the Few-shot element. We quantitatively compare the navigation sentences in the datasets created by each prompting method. Additionally, by examining the trends in the generated text for each prompt method, we investigate the impact of the elements that make up HLTFC.

As an evaluation method, we apply our proposed G-ACF to each dataset and conduct a multi-perspective evaluation. Since this experiment focuses on assessing the dataset quality, traditional text-based evaluation metrics, which require ground truth data, are not applicable. The navigation sentences in HLG need to evaluate the relationships between various elements of the images and texts. Therefore, even metrics like the CLIPScore, commonly used in Vision & Language tasks, cannot provide an adequate evaluation. On the other hand, it is believed that it will be possible to evaluate methods using GPT-4o, which has multimodal reasoning capabilities and extensive knowledge, by designing appropriate prompts. However, since we use custom prompts, the reliability of automatic evaluation by LLMs still requires verification. Therefore, by conducting multifaceted evaluations using G-ACF, we ensure reliability.

The quantitative evaluation results are shown in 1. From Table 1, we confirmed that our method achieved the highest accuracy, followed by CoT, and Few-shot in descending order of accuracy. In addition, the correlation between the two evaluation models is weak for all prompts, indicating that they are evaluating from different perspectives. Next, we show some of the occurrence probabilities of the words indicating the intersection in the sentences in the data set in Table 2. From Table 2, we observed that while Few-shot and HLTFC exhibited a wide spread in occurrence probabilities, CoT showed less dispersion. In Few-shot, the occurrence probability was high even for cases other than the examples presented in the prompt. In CoT, there was a tendency to frequently use the intersection expression ”where the [Object] is” throughout the entire dataset. These findings suggest that Few-shot prompts contribute to more flexible sentence expressions, while CoT promotes higher-quality sentence generation. Furthermore, it can be said that HLTFC generated a dataset with rich sentence diversity and expressive power.

4.2 Effectiveness Studies of G-ACF

We conducted experiments to investigate the effectiveness of our proposed GPT guided Auto Cleaning Framework (G-ACF). Using G-ACF, we cleaned the dataset created by HLTFC and train the VLM model. The generation of Navigation sentences requires general knowledge of driving scenes, such as road structures. Furthermore, the ultimate goal of HLG is in-vehicle implementation. Therefore, in this experiment, we performed fine-tuning using LLaVA, an open-source state-of-the-art Vision & Language


Evaluation Criteria Prompt Score: 9.7
Human thought process prompt Score: 9.3
G-ACF Score: 9.5
Evaluation Criteria Prompt Score: 8.0
Human thought process prompt Score: 5.0
G-ACF Score: 6.5

Guidance sentence:
Turn left at the intersection where **the red car** is heading.

(a) Left turn scene

Guidance sentence: Follow **the red car** to turn right at the intersection.

(b) Right turn scene

Figure 4: Qualitative comparison results. By combining the two evaluation models of G-ACF, it is possible to perform appropriate evaluation.

Large Model (VLLM). For comparison, we altered the data filtering methods in G-ACF and compared the accuracy after fine-tuning each dataset. This is because no suitable data cleaning method currently exists for proper comparison in HLG tasks.

The three filtering methods are as follows:

- No filtering applied.
- Threshold set to 8 points, taking the logical OR of the two evaluate models
- Threshold set to 8 points, taking the logical AND of the two evaluate models

The quantitative evaluation results are shown in Table 3. From Table 3, we confirmed that filtering using logical OR and logical AND is highly accurate.

As the qualitative comparison, we show the generated sentence from the trained LLaVA model in Figure 4. In this example, we show the results of the model trained on the OR filtered data, which had the highest accuracy. As in previous experiments, we adopt the G-ACF method we have proposed to evaluate the output sentences from multiple perspectives. Figure 4(a) presents an example where both evaluation models judged the guidance sentence to be highly accurate. The generated navigation sentence which references a "the red car" traveling ahead in the same direction, is intuitive and appropriate for the driver, confirming the correct evaluation. Figure 5(b) shows an example where one evaluation model rated the sentence highly, while the other rated it less accurate. The sentence used the verb "follow" in relation to a vehicle that had stopped and was traveling in the opposite direction. This is clearly an inappropriate instruction. This is a sample that has been evaluated correctly through a multi-perspective evaluation using two evaluation models. From these results, we can say that the cooperative evaluation method using multiple LLM in the G-ACF that we proposed is effective.

Table 3: Accuracy comparison of the fine-tuned model using datasets filtered by different methods. The G-ACF values represent the average scores from the two evaluation models.

	Evaluation Critica prompt	Human thought prompt	G-ACF
No fillterd data	8.96	8.92	8.94
OR fillterd data	**9.18**	**9.29**	**9.24**
AND fillterd data	9.11	9.18	9.14

5 CONCLUSIONS

In this research, we propose a method to automatically create a high-quality navigation sentence dataset solely from image data, aiming to realize HLG. First, we utilized GPT-4o to automatically generate navigation sentence data through our proposed prompting method. Experiments confirmed the effectiveness of our prompting method and the impact of each element within the prompt. Next, we constructed a data evaluation framework in the MoE format to automatically clean the generated data. The experiments showed that our proposed method is capable of providing appropriate evaluations even for complex tasks such as HLG. In the future, it will be necessary to focus on lightweighting and speeding up for implementation in real-world vehicle environments.

REFERENCES

(2024). Vdc: Versatile data cleanser based on visual-linguistic inconsistency by multimodal large language models. *International Conference on Learning Representations (ICLR)*.

Achiam, J., Adler, S., and et al (2023). Gpt-4 technical report. *Preprint*.

Allen, G. L. (1999). Cognitive abilities in the service of wayfinding: A functional approach. *Professional Geographer*.

Banerjee, S. and Lavie, A. (2007). Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *Proceedings of the Second Workshop on Statistical Machine Translation*.

Barrow, K. (1991). Human factors issues surrounding the implementation of in-vehicle navigation and information systems. *SAE Transactions*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems 33 (NeurIPS)*.

Burnett, G. (2000). 'turn right at the traffic lights':the requirement for landmarks in vehicle navigation systems. *The Journal of Navigation*.

Chmielewski, M. and Kucker, S. C. (2019). An mturk crisis? shifts in data quality and the impact on study results. *Social Psychological and Personality Science*.

Evans, G. W., Skorpanich, M. A., Gärling, T., Bryant, K. J., and Bresolin, B. (1984). The effects of pathway configuration, landmarks and stress on environmental cognition. *Journal of Environmental Psycholog*.

Fu, J., Ng, S.-K., Jiang, Z., and Liu, P. (2024). Gptscore: Evaluate as you desire. *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

He, X., Lin, Z., Gong, Y., Jin, A.-L., Zhang, H., Lin, C., Jiao, J., Yiu, S. M., Duan, N., and Chen, W. (2024). Annollm: Making large language models to be better crowdsourced annotators. *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. (2021). Clipscore: A reference-free evaluation metric for image captioning. *Empirical Methods in Natural Language Processing (EMNLP)*.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *In Proceedings of the Workshop on Text Summarization Branches Out (WAS)*.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023a). Visual instruction tuning. *Advances in Neural Information Processing Systems 36 (NeurIPS)*.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023b). G-eval: Nlg evaluation using gpt-4 with better human alignment. *Empirical Methods in Natural Language Processing (EMNLP)*.

Madhyastha, P., Wang, J., and Specia, L. (2019). Vifidel: Evaluating the visual fidelity of image descriptions. *Association for Computational Linguistics (ACL)*.

Nambata, M., Shimomura, K., Hirakawa, T., Yamashita, T., and Fujiyoshi, H. (2023). Human-like guidance with gaze estimation and classification-based text generation. *International Conference on Intelligent Transportation Systems (ITSC)*.

Oh, S., Lee, S. A., and Jung, W. (2023). Data augmentation for neural machine translation using generative language model. *arXiv:2307.16833*.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. *Annual Meeting of the Association for Computational Linguistics (ACL)*.

Passini, R. (1984). Spatial representations, a wayfinding perspective. *Journal of environmental psychology*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *arXiv:2103.00020*.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training. *Preprint*.

Tom, A. and Denis, M. (2003). Referring to landmark or street information in routedirections: What difference does it make? *COSIT 2003 Lecture Notes in Computer Science 2825*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems 30 (NIPS)*.

Wang, J., Meng, L., Weng, Z., He, B., Wu, Z., and Jiang, Y.-G. (2023). To see is to believe: Prompting gpt-4v for better visual instruction tuning. *arXiv:2311.07574*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems 35 (NeurIPS)*.

Xu, H., Xie, S., Huang, P.-Y., Yu, L., Howes, R., Ghosh, G., Zettlemoyer, L., and Feichtenhofer, C. (2023). Cit: Curation in training for effective vision-language data. *International Conference on Computer Vision (ICCV)*.

Yu, Y., Zhuang, Y., Zhang, J., Meng, Y., Ratner, A., Krishna, R., Shen, J., and Zhang, C. (2023). Large language model as attributed training data generator: A tale of diversity and bias. *Advances in Neural Information Processing Systems 36 (NeurIPS)*.

Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., and Hu, X. (2023). Data-centric ai: Perspectives and challenges. *In Proceedings of the 2023 SIAM International Conference on Data Mining(SDM)*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. *International Conference on Learning Representations (ICLR)*.

Zhao, W., Peyrard, M., Liu, F., Gao, Y., Meyer, C. M., and Eger, S. (2019). Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. *Empirical Methods in Natural Language Processing (EMNLP)*.