Spatial Referencing for Large Language Models in Automotive Navigation Tasks

Khanh Huynh
BMW Group
Munich, Germany
LMU Munich
Munich, Germany
khanh.huynh@bmw.de

Jeremy Dillmann
BMW Group
Munich, Germany
jeremy.dillmann@bmw.de

Sven Mayer
TU Dortmund University
Dortmund, Germany
Research Center Trustworthy Data Science and Security
Dortmund, Germany
info@sven-mayer.com

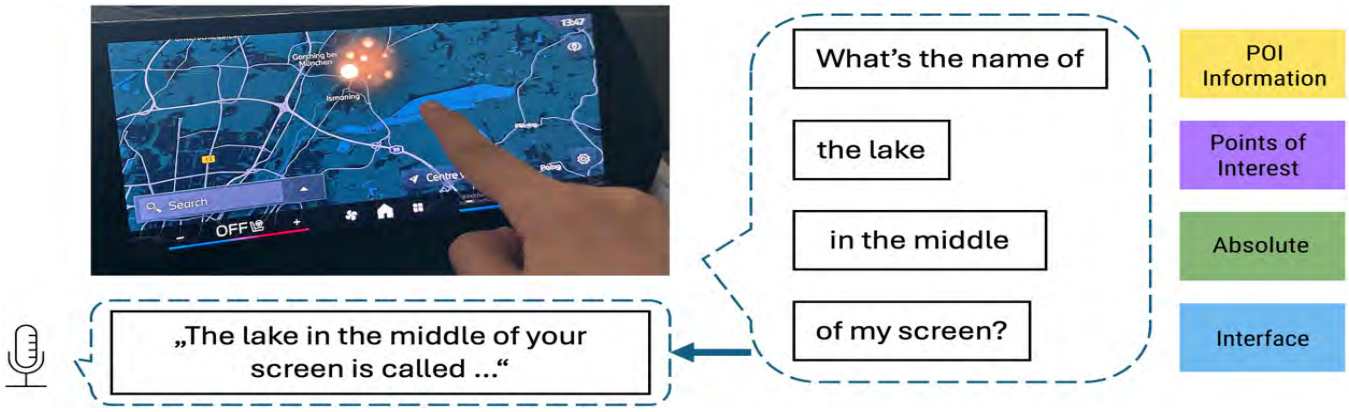


Figure 1: Example of spatial referencing in in-car navigation: user asks "What's the name of the lake in the middle of my screen". System replies by grounding to the display screenshot. Categories of taxonomy (POI information, Points of Interest, Absolute, Interface) can be mapped to the user utterance.

Abstract

In human-human conversations, a shared visual layer allows conversation partners to refer to visual elements through spatial references - such as "on the left" or "the blue pen next to you". Current voice user interfaces, however, lack the context needed to interpret such references, limiting their naturalness. This capability is particularly valuable for in-car interactions, where combining voice and graphical interfaces offers opportunities for more fluent and effective interaction while driving. In this work, we integrate a multimodal large language model for an in-car infotainment system to enable the interpretation of spatial references. Through a user study (N=21), we collect and analyze user utterances to investigate within the context of automotive navigation tasks. As a result, we created a taxonomy that categorizes diverse strategies participants used to reference on-screen elements. Our findings contribute a framework for understanding spatial referencing behavior in vehicles and inform the design of future multimodal in-car systems.

MUM '25, Enna, Italy

ACM ISBN 979-8-4007-2015-4/25/12
https://doi.org/10.1145/3771882.3771917

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**.

Keywords

Conversational Agents, Multimodal Large Language Models, Human-Vehicle Interaction, Multimodal Interaction, Automotive Navigation

ACM Reference Format:
Khanh Huynh, Jeremy Dillmann, and Sven Mayer. 2025. Spatial Referencing for Large Language Models in Automotive Navigation Tasks. In *24th International Conference on Mobile and Ubiquitous Multimedia (MUM '25), December 01–04, 2025, Enna, Italy.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3771882.3771917

1 Introduction

In-vehicle interaction often involves multiple interfaces simultaneously, such as the built-in infotainment system, the steering wheel, and haptic buttons [27]. The primary interaction modality for infotainment functions is usually a graphical user interface (GUI). For example, navigation systems provide drivers with a visual representation of routes and points of interests (POIs) inside maps and navigation applications. While GUIs are a core part of automotive interfaces, voice user interfaces (VUIs) offer advantages with a hands-free nature and lower visual distraction [23, 26]. Recent advances, particularly the integration of large language models

(LLMs) in VUIs, present an opportunity to make conversations more natural and human-like [20]. One decisive aspect in human-human conversation is our ability to reference spatially in a shared visual layer. For example, when we are standing next to each other, we use verbal spatial referencing to speak freely about the location of objects relative to us, like what is to our "left" or "right." Other examples could be using shape descriptions like "heart-shaped" or colors like "green." Spatial referencing, applied to the GUI inside the vehicle, allows the driver to reference visual elements on the GUI using natural language, such as "pick the route that is highlighted in blue" or "take the marker on the bottom right." Although the LLM's capacity to process natural language appears to make this form of communication feasible in the vehicle [29], little is known about how and why users would use spatial referencing and what kind of technical solutions are necessary. Further research is needed to understand user interaction patterns before identifying solutions that can best support a seamless user experience.

In recent research, the modality of including images as contextual input in LLMs was introduced and explored as multimodal large language models (MLLMs) [12, 34, 38]. This modality has been used in visual question answering (VQA) of users asking general questions about pictures [1, 11]. Possible questions are mostly generalized, like "What's the name of the building?" or "Which dog breed is this?", focusing on the big picture of the image. Despite these advancements, visual MLLMs still encounter difficulties generating correct outputs with a tendency to produce erroneous or hallucinated responses [10, 18]. Tackling these challenges, research investigating spatial referencing is largely resolved around the improvement of spatial reasoning, which describes the ability of a system to link spatial expressions (e.g., "in front of", "to the left of", "above") to concrete visual elements. In comparative evaluations, humans achieve accuracies above 95%, while state-of-the-art MLLMs only reach around 70% accuracy on benchmarked spatial reasoning tasks [19]. Spatial referencing has also been investigated in human-robot interaction. Li et al. [17]'s findings highlight how ambiguity can reduce clarity and accuracy of natural language. Instructions that were perspective-independent (e.g., "the yellow block in the middle") were much easier to follow than those relying on unspecified viewpoints (e.g., "the block to my left"), which often led to misinterpretations. Within the domain of GUI understanding, FerretUI [39], for instance, provides an MLLM enhanced with the ability to ground and reason about on-screen elements. By dividing GUIs into sub-images for separate encoding and fine-tuning the model on a curated dataset of questions about the GUI. Further advances in grounding frameworks and improving MLLM performances have led to state-of-the-art systems for GUI understanding [2, 9, 37, 39]. While these methods advance GUI understanding, they are domain-specific and remain limited to web, desktop, or mobile settings. VUIs rely on speech, which also introduces unique challenges for spatial referencing, since users must describe GUI elements verbally. In automotive contexts, visual content is dominated by navigation systems with maps and route information, which provide a particularly rich source of graphical context. However, empirical research appears to be lacking on how users interact with such systems in real time and investigates how drivers would spatially reference in in-car GUIs. To address this gap, we implement a prototype in-car VUI that integrates MLLMs with image inputs to enable spatial referencing within the navigation context. We conducted a user study to collect and analyze how drivers spatially reference in-car GUIs during navigation tasks. Our contributions are: (1) the design and implementation of a spatial-referencing-enabled in-car VUI, and (2) empirical insights into the challenges and opportunities of spatial referencing in automotive contexts, derived from a controlled user study conducted in a stationary vehicle.

Using the results of the study, we used thematic analysis to create a taxonomy that represents how users use spatial referencing to speak about displayed content on the screen in the navigation context. We further analyzed the frequency and sequence of different reference types. The implementation of spatial referencing presents a step forward in creating a hands-free and intuitive VUI and GUI experience inside the vehicle. At the same time, the taxonomy will provide an understanding of user behavior that will help with the continuous design and implementation of human-like VUIs.

2 Related Work

> "There are classes of things that are done better with speech and natural language than with direct manipulation...And when speech and language interfaces become more conversational, they will take their place along with direct manipulation in the interface." [7]

Already in the 1990s, Don et al. [7] envisioned speech as a natural complement to GUIs, suggesting that conversational VUIs would one day blend seamlessly with direct manipulation. One aspect of a blend of VUI and GUI is the ability to reference what you see in the real world, thus enabling spatial referencing. We first describe prior research on spatial references in conversational speech interaction and their role in in-car systems. We then dive into the topic of how such spatial references could be enabled and what has been done in that domain.

2.1 Interactions through combining VUI and GUI and the Automotive Context

With the launch of commercial VUIs, speech interaction has gained popularity. Since then, speech interactions have become more well-known, and an increasing number of assistants are being released by tech giants, trying to improve the experience and find use cases for speech assistants. However, a study by Mahmood et al. [20] revealed that the most common uses of VUIs still include system-based commands like asking for the weather (70% of the participants) and setting reminders, timers, and alarms (65% of the participants). Thus showing that system-based commands like "What's the weather?", "Set timer to 10 minutes." and "Set a reminder for 2 PM." are still most known to users. This raises the question, even if technology is moving forward, when will speech interaction take the shift to a more human-like conversation? When referring to VUI design guidelines, a meta-analysis, Murad et al. [22], found that one of the guidelines was to design conversational interaction that maps to real-world conversational norms and dialogue patterns. VUI interactions should align with the user's mental model, and part of this could also be applied to the information from a shared visual layer. Research combining speech with gaze cues demonstrates the benefits of multimodality for spatial referencing. When participants

could see their partner's gaze position in addition to hearing speech, they completed tasks more efficiently and with greater precision [24]. This suggests that speech alone may be insufficient in contexts requiring spatial grounding, such as in-vehicle interaction, where shared visual references could reduce ambiguity. When it comes to in-vehicle visualization, more and more car functions are being displayed in the digital realm. Automotive is experiencing a shift from traditional physical buttons to GUIs in car interiors. Switching buttons, knobs, and switches with interactive displays [5, 21, 27]. In the specific use case of navigation, the GUI comes in as a great way to present the map and information about routes, POI, and occurring route incidents. A focus in the automotive context is the navigation use case, with the challenges of the most visually and graphically displayed information. The role of a VUI is hands-free use and requires the least eye glances, but at the same time, it requires a cognitive load to process the information [26]. Most of the time, the information is not even presented in any visual way. If the GUI visually presents the information, it requires drivers' attention and the most eye glances. Integrating speech and visual references could reduce these trade-offs, creating more natural, efficient in-vehicle interactions. Spatial terms in speech can serve as a bridge between what users see and what they say, enabling conversations about shared visual content. However, spatial references are inherently ambiguous, as they can be interpreted egocentrically (relative to the speaker) or allocentrically (relative to the conversational partner's perspective) [17]. Understanding these nuances is key to effective multimodal system design.

2.2 Spatial Reasoning for Multimodal Large Language Models

The technical evolution from LLMs to MLLMs marks a significant step forward in the work and research of LLMs. MLLMs are models that include reasoning not only on text but also on audio and images. In work focusing on GUI interpretation, FerretUI used an MLLM to enhance the understanding of mobile screens. FerretUI specifically addresses the challenges of processing mobile screens, where elements like icons and text are often small and detailed. Improvements in visual detail were handled by dividing the GUI into sub-images for separate encoding [39]. Training the model on a curated dataset of elementary and advanced questions about the GUI further enhances its ability to perform advanced tasks like "Where can I find the app store?". This approach surpassed GPT-4V on all the elementary UI tasks.

However, integrating the vision modality also comes with limitations. It has been demonstrated that MLLMs tend to provide responses that are inconsistent with real-world knowledge or user inputs, which are known as hallucinations [10]. Hallucinations with vision-based MLLMs are described as image content answering that is inconsistent with the image content itself. MLLMs have been known to suffer from this phenomenon.

Recent work also implies that MLLMs do not generate output heavily based on visual information when textual context is provided. When given both visual and text data, MLLMs tend to rely more on the text. When given mismatched visual and correct text data, the performance is not necessarily hindered, implying even more that MLLMs are not heavily reliant on visual context, especially when textual clues are provided. Not only that, but the absence of visual input even leads to a better accuracy across all questions, showing that visual inputs might even hinder the accuracy [32, 41]. This contrasts with human capabilities, where visual cues might aid in understanding.

Spatial reasoning is an especially difficult area. Studies show that MLLMs struggle with basic relations such as "left of" or "right of," rarely exceeding 60% accuracy even with extensive training data [3, 13, 19]. In a study by Liu et al. [19], most visual-based MLLMs struggled to exceed 60% accuracy even with more training examples. Training objectives partly explain this gap: contrastive models perform somewhat better than generative ones, but both remain weak at handling spatial terms [13]. A further obstacle lies in pre-training data. Even if this is the case, prepositions are rarely needed to make the model perform well on the contrastive training objective [13, 40]. It is also noted that pre-training models use large datasets like LAION, which was also used to train OpenCLIP [28]. In LAION, prepositions like "under" or "left of" only occur 0.2% of the time [13]. When prepositions are used, they can be ambiguous due to the viewer's perspective or vary in interpretation of the same preposition [3, 13]. "In front of" could mean close to the viewer of the image or ahead of the elements that are portrayed in the image. For example, it might be shown two images or two pieces of text and asked to determine which is most relevant to the task. These models also rely on the large batch size to differentiate similar examples, but not from prepositions. For example, distinguishing "Golden retriever" from other dog breeds. Some prepositions are much more common than others. For example, "dog under the table" vs. "dog on the table." These limitations matter directly for multimodal VUI GUI systems in cars, where spatial referencing is essential for grounding speech in shared visual context. Stappen et al. [29] proposed one of the first applications of MLLMs in vehicles, focusing on diagnosing technical issues by combining verbal prompts with visual inspection. Such multimodal approaches could yield faster and more personalized solutions, but only if spatial reasoning capabilities are strengthened.

3 Research Questions

We investigate how to enable spatial referencing for LLM-based VUIs in the car and investigate user behavior with such interaction through the following research questions:

Research Question 1. While prior work has demonstrated multimodal LLM capabilities, such as VQA, research has largely focused on general scene understanding (e.g., "Which dog breed is this?", "What do I see in the picture?") rather than the interpretation of spatial references [1, 11]. Existing research emphasizes the limitations of spatial reasoning, yet there is limited understanding of how such models can be adapted to handle spatial references in task-specific contexts [10, 18]. Therefore, we pose our first research question: **RQ1: How can LLM-based VUIs be designed to understand user utterances containing spatial references?**

Research Question 2. Prior research in the automotive domain has not yet fully explored spatial referencing for automotive GUIs. Navigation applications, for example, are highly visual, provide

interactive maps, and are spatially structured. Enabling this novel type of interaction raises questions about how users will interact with the system, which strategies will be impaired to express spatial references, and in which situations they will be motivated to do so. Speech enables another complex challenge that provides a fast and dynamic interaction modality [22, 26]. Understanding user behavior will help address the challenges in enabling this type of interaction in LLM-based VUIs. Accordingly, we address the following second research question: **RQ2: How do users structure and utter spatial references during interactions with a VUI in in-vehicle navigation tasks?**

4 Enabling Spatial Referencing for a Voice User Interface in the Car

To address **RQ1**, we implemented a prototype system that allows an LLM-based VUI to process spatial references in the context of an automotive infotainment system, specifically the navigation application, as shown in Figure 2. Our approach combines the multimodality of LLMs, which refers to their ability to process and integrate different data modalities, such as text, images, and structured data, with LLM function calling. In this case, screenshots from the central display in the vehicle were integrated as images. In this approach, we define understanding of spatial references as the ability of the system to:

(1) Ground spatial expressions (e.g. "at the left top corner", "blue colored",...) in GUI elements
(2) Generate contextually appropriate responses that align with the user's intent

The underlying system builds on an existing in-car infotainment platform connected to a cloud-based backend. The backend hosts the LLM, with GPT-4o as the model in use, and handles LLM requests. For the speech component, the speech input was processed using Alexa's speech-to-text service, and then the processed text was sent to the LLM backend. Similarly, the speech output was using the text-to-speech services to output the text with the Alexa voice. For the image modality, screenshots of the navigation application as displayed on the screen were continuously captured and sent to the backend in PNG format. Screenshots were preprocessed to an optimized resolution of 800 x 400 pixels for reliable recognition of GUI text and elements. Pilot testing ensured that map text and interface elements (e.g., icons) were interpreted without errors in recognizing the elements. Lower resolutions led to recognition errors, such as misread texts or responses indicating an inability to interpret the image (e.g., "The image is a bit blurry, but..."). In addition to the screenshot, structured metadata about displayed GUI elements was supplied in a JSON format. For example, the metadata specified domain-specific elements such as `"blue_highlighted_route": "current_route"` or `"red_lines": "traffic_congestion"`. This ensured that the model's reasoning was grounded not only in its general world knowledge but also in the specific semantics of the automotive-specific GUI. It is specified that other visual elements within the car, such as ambient lights, are not part of the GUI to avoid potential conflicts with function calling, as they are also shown in colors. The conversation context combined four elements: (1) conversational history, (2) the user's utterance, (3) the screenshot of the display,

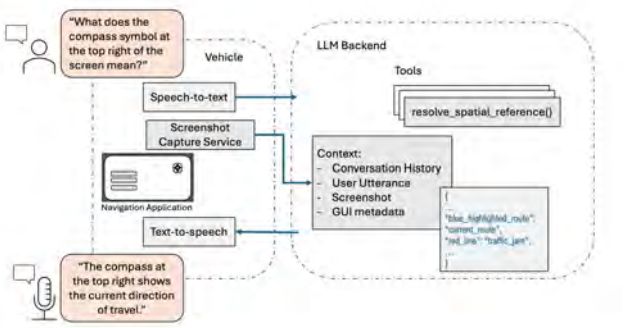


Figure 2: System architecture of the LLM-based voice user interface (VUI) for spatial referencing in automotive navigation. Speech is processed via speech-to-text and text-to-speech. Screenshot capture service sends images to the backend. Screenshot capture service sends images to the backend. LLM Backend (GPT-4o) integrates conversation history, user utterance, screenshots, and GUI metadata, and invokes the resolve_spatial_reference() tool to ground expressions to specific GUI elements (in this example, the compass symbol).

and (4) structured metadata about the GUI. Spatial referencing was enabled through LLM function calling. Function calling extracts structured information and matches a pre-defined schema, then includes the response in the context of the conversation. When the user referred to a visual element (e.g., "the lake above", "on the left", "the middle icon"), the system invoked a function call named resolve_spatial_reference(), which structured the model's output according to a predefined schema. The tool then appended the structured output, together with the GUI metadata and the screenshot, to the ongoing conversation history, allowing the LLM to reason about spatial relations in subsequent turns. By combining image input through screenshots, structured metadata, and function calling, our implementation demonstrates a practical approach to enabling spatial referencing in LLM-based VUIs for automotive navigation.

5 Understanding Spatial References in Automotive Navigation Tasks

We conducted a study in a standing vehicle, using tasks in the context of navigation. With this study, we collect user utterances and create a taxonomy based on the user utterances that were spatial references.

5.1 Procedure

We designed the study in three parts. The first part included formalities and an introduction to the topic. Formalities also meant filling out consent forms and a demographic questionnaire. We introduced the users by letting them have the first interaction with the system with a small task. The introduction task was to inform themselves about a POI and navigate to it.

For the second part, we introduced three main tasks (see Section 5.2) to the participants, which will be further explained in the

following section. Participants interacted with the system to complete the three tasks. At the beginning of each task, we introduced the scenario verbally (e.g., "Imagine this scenario..."), explained the specific goal, and clarified that the task would conclude once the goal was achieved. We confirmed participants' understanding before starting each task. Depending on the participant's interaction pace and the task complexity, each session lasted approximately five to ten minutes.

The third and last part was an ending questionnaire that concluded the study with general questions and evaluations. The study was carried out in German. For a speech study, using participants' mother tongue yields results that better reflect natural interaction, as language proficiency does not add an additional barrier to cognitive load [14, 31]. We translated all participants' answers into English when quoted.

5.2 Tasks

We selected three tasks that differed in their goals and levels of interaction within the navigation system to ensure diversity in user engagement. This selection allowed us to capture participants' underlying motivations across varying interaction contexts. We selected the three tasks to cover distinct interaction goals: a goal-oriented and information-seeking task (POI Search), a spatial task involving non-GUI elements and free map exploration (Map Navigation), and an exploratory and GUI-focused task without predefined objectives (GUI Exploration), ensuring a comprehensive examination of user interaction behaviors across functional, spatial, and exploratory contexts. The three authors selected the tasks through an iterative design process to ensure that they aligned with the overall study goals.

5.2.1 Point of Interest Search. In this task, we asked participants to search for a charging station. The only condition was to search for one with at least 200 Kilowatt hours (kWh). The task included asking for information about the charging station before choosing one. There was no condition about which information exactly. Examples like opening hours, rating, and available spots were given, but the participant had the option to keep their personal motivations in mind. The task would end after the participants had chosen a charging station and started the navigation towards it. The purpose of this task was to include one of the most important use cases within the navigation system, the POI search. The task allowed us to investigate how often and in what ways users reference when weighing options with POIs.

5.2.2 Multimodal Interaction in Map Navigation. In this task, we showed participants a map element on a piece of paper. Their goal was first to identify the element on the map, then obtain information about it, and finally start the navigation to it. The task was free in the interaction of using touch to locate the map to the lake. Participants could freely use speech or touch to complete the task. To avoid bias, the map element was consistently introduced as a neutral "element on the map" without descriptive hints such as "lake". With the automotive design choice, the map is very minimal in its information. Most street names and other names of POIs are not shown. This is unlike the usual GUIs inside of known navigation applications like Google Maps or Waze Navigation. This task examined how users combined GUI and VUI interaction, and how spatial references emerged when participants located and described map elements.

5.2.3 User Exploration of Graphical User Interface. The last task was an exploration task to find out information about the navigation application, focusing on the GUI only. Participants were encouraged to ask about interface elements, icons, and map components. The task concluded either after participants felt they had no further questions or after five minutes had elapsed. The purpose of this task was to observe spontaneous spatial referencing strategies and to better understand participants' motivations for referencing visual content in a less structured and goal-oriented setting.

5.3 Apparatus

We conducted the study in a standing vehicle, meaning driving-related interfaces were not accessible during the study. This setup was chosen as an exploration to capture the full range of user behaviors in navigation tasks without the cognitive load of driving. Investigating navigation interactions in a standing vehicle also reflects a relevant use case, since such tasks are often initiated while the vehicle is stationary (e.g., before departure or during breaks). The participants were sitting in the driver's seat. We did not use a wake word, but used the haptic microphone button at the steering wheel for the VUI activation (see Figure 3a).

5.4 Participants

We recruited 21 participants (8 female and 13 male) from BMW Group to take part in the study. The age range was 23 to 43 years (M = 27.9, SD = 4.7). Although we did not assess driving experience, participants' employment at BMW Group suggests a high level of familiarity with in-vehicle systems and driving-related contexts. For experience with general speech assistant participants, answers ranged from 1 to 4 (M = 3.05, SD = 1.00, on a scale of 0 to 5, where 0 = unknown and 5 = used regularly). The experience with the automotive-specific speech assistant ranged from 0 to 3 (M = 1.38, SD = 1.00 on a scale of 0 to 5, where 0 = unknown and 5 = used regularly). The affinity for technology interaction (ATI) questionnaire was used to measure the participants' tendency to engage or avoid interactions with new technologies [8, 16]. The ATI score ranged from 30 to 45 (M = 38.80, SD = 3.63), with 45 being the maximum score of the scale.

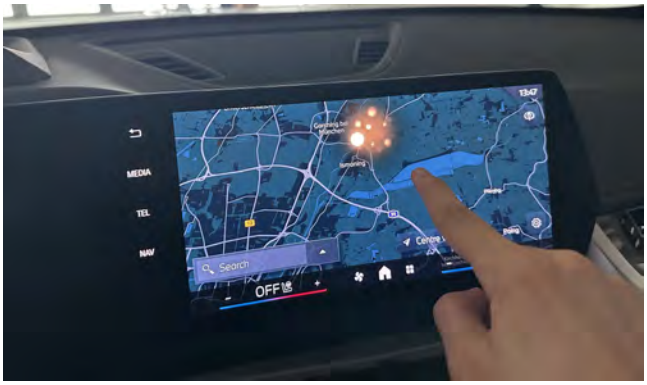
6 Results

The following section will go into the system performance of our implementation to specify the challenges that occur later. Then dive deeper into the taxonomy resulting from the user study. In total, we collected 743 utterances in 63 runs. Out of these, we categorized 116 (15.61%) utterances as spatial references. The 116 spatial references were examined through thematic analysis, as described in Section 6.2.

6.1 Prototype Performance

Within the 116 spatial references collected in the study, we analyzed system responses based on correctness and tool usage. Each response was labeled as either a correct or incorrect interpretation

(a) Study setup from the participant's view with the microphone button on the right side of the steering wheel and the display with the map in the middle.



(b) Alternative perspective showing the on-screen interaction on the central display.

Figure 3: Study apparatus from two perspectives: participant's seating view (a) and display-focused interaction view (b).

of the visual context, and tool invocation was additionally checked for whether it was correctly triggered. 72 out of 116 spatial references were successfully interpreted. 17 times the responses were misinterpreted, even though the tool was correctly invoked. In 23 cases (out of 743 utterances), the tool was called unnecessarily, for example, when the user asked about a color unrelated to the display. In 73 cases, the tool was not called, although it would have been appropriate. Most of the cases occurred because a screenshot was already in the context, leading the model to skip another tool call. However, this can lead to mismatch as the screen content might change within one run. In other cases where the tool was not called, the LLM simply agreed with the user instead of grounding through the tool. For example, when asked: *"But at the top right, that's the current time showing on the screen right?"*, the model responded: *"You are right, at the top right of the screen you can see the current time!"*. The different counts (e.g., correct responses, incorrect tool calls, or missed tool opportunities) represent overlapping classifications: even when a response was correct, the grounding tool might not have been invoked, meaning the visual context was not updated. Such cases could lead to inconsistencies if the on-screen

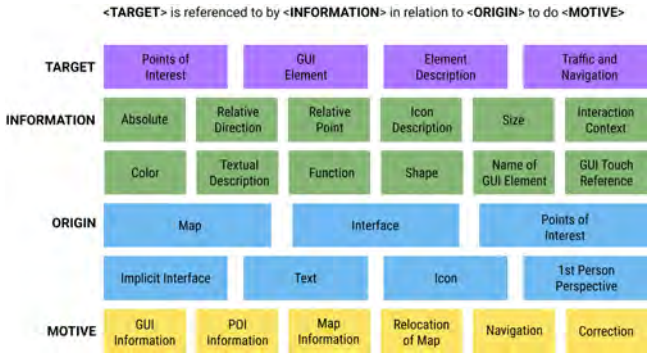


Figure 4: Taxonomy of utterances scenario. A spatial reference can be built using the structure <Target> is referenced to by <Information> in relation to <Origin> to do <Motive>.

content changed, which is why the sum of these instances exceeds 116.

This analysis led to the following distribution:

- **72 correct vision responses**, the system successfully interpreted the spatial reference and provided the correct answer.
- **17 incorrect vision responses**, the system misinterpreted the spatial reference, despite invoking the tool.
- **23 incorrect tool calls**, cases where the system invoked an unnecessary tool or selected the wrong one.
- **73 missed tool opportunities**, cases where a tool should have been invoked but was not.

Overall, this breakdown shows that while the system was able to produce correct responses in more than half of the cases (56.03%), errors frequently stemmed from unreliable tool invocation. This highlights that the core challenge lies less in generating a correct response once the appropriate tool is used, and more in the model's decision of *when* and *how* to invoke tool calls for spatial reasoning.

6.2 Utterance Taxonomy for Spatial References

To analyze the user behavior, we looked at each user utterance that was a spatial reference and used thematic analysis to make sense of the data [4, 30]. We used the tool Atlas.ti to code them. For this, two researchers independently coded 20% of the study runs of randomly picked participants. After this, a third researcher joined the discussion of the codes. We refined the codes and formed a joint code book. We repeated this process and measured the inter-rater reliability using Krippendorf's alpha [15]. After resulting in Krippendorf's alpha of 0.84, one researcher coded the rest of the study runs. The codes provided insights into the varying ways of spatial references and resulted in a taxonomy. The utterance scenario taxonomy [33] was created to demonstrate the themes relevant to building a spatial reference utterance. The themes are Target, Information, Origin, and Motive. A spatial reference can, therefore, be built using the following utterance structure:

<Target> is referenced to by <Information> in relation to <Origin> to do <Motive>

The taxonomy captures both the linguistic strategies and the underlying intentions behind participants' spatial references. By filling the placeholders with one or multiple types of tokens, every spatial reference of the study can be mapped fully (see Figure 4), regardless of variations of phrasing. Table 1 provides an overview of all categories within each theme, along with example utterances and frequency of occurrence. To make the understanding easier, we selected a representative utterance from the study to further explain each theme in more detail: *"I am looking for the name of this large, long lake at the bottom of the screen."* In the following, the themes will be described in more detail using the example utterance.

Target illustrates the element or entity being referenced in the display, in this case, the *"lake"*. It indicates which visual component the user's spatial reference is directed towards. When combined with the other themes, identifying the Target is crucial for interpreting the exact element the user is referring to.

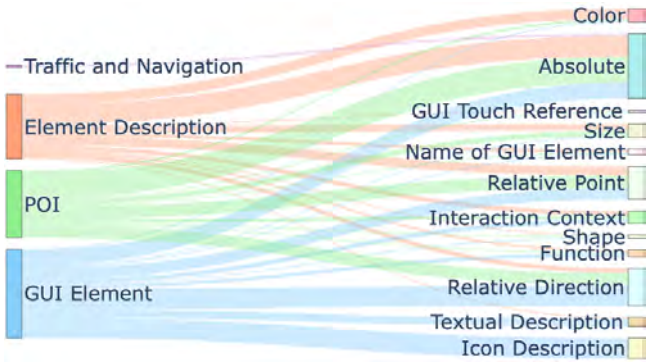
Information describes the way in which the spatial reference was formulated and what type of detail comes with it. This information will then be used to pinpoint the exact element that was spatially referenced to. In the example, the lake is described by its size (*"large"*), shape (*"long"*), and its relative position (*"at the bottom of..."*) A single utterance can combine multiple Information categories to describe the visual element in detail. Thus, the Information theme answers the question of how exactly the user references visual elements.

The **Origin** theme specifies the reference point or frame to which the Target element is described. In other words, it captures what the user relates the Target to. In the example, the lake's position is anchored in relation to the interface (*"at the bottom of the screen."*). Other possible Origins include the map itself, the overall interface, specific elements like text, icons, or POIs within the interface, or even the user's own perspective. It is noted that there can also be no relation worded, which results in the category "implicit interface." This is due to the tasks revolving around the interface and, therefore, the assumption that the interface is always the Origin if not otherwise mentioned.

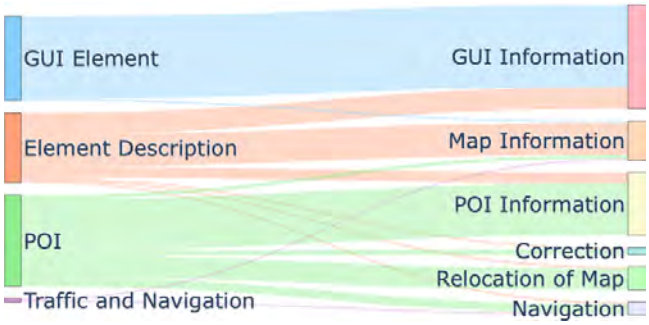
Motive explains the underlying intent behind the spatial reference or what the participants want to achieve with it. Here, the participant seeks POI information (*"I'm looking for the name..."*) and expects an answer in that direction. By capturing the underlying intention, Motive explains not only what was said, but why it was said.

6.3 Analyzing Relationships and Frequencies of Spatial Reference Categories

The first frequency and the relation between the themes Target and Information is shown in Figure 5a, with element description and POI being the most prominent. Element description shows connections to all Information categories, with particularly strong links to absolute references, color, relative points, and size. POIs are most frequently described using absolute references, relative directions, and relative points. The use of absolute references, relative directions, and relative points across various Target categories with high links underscores the importance of spatial communication, indicating a frequency of references in relation to other POIs



(a) Frequencies and relationships where Target is referenced by Information.



(b) Frequencies and relationships where Target is referenced by Motive.

Figure 5: Frequencies and relationships of the two themes. (a) Target referenced by Information. (b) Target referenced by Motive.

or GUI elements. When looking at the Information icon description and textual description, it is noticed that it is almost only in relation to GUI elements, indicating that other Target categories are rarely being described by icons or textual descriptions.

As seen in Figure 6, the relations and frequencies of the <Origin> in relation to <Information> are shown. With absolute in the Information theme being the most frequently used, it draws mostly to the interface as a whole and the map. The system as a whole is often mentioned as *"on the navigation system"* or *"on the display"*. For the map, the same is applied, but mentioning the map instead. *"On the map"* is by far the most mentioned, with 28 quotations within 116 spatial references (24.14%). Map and Interface are the most prominent categories for the Origin theme. The frequency of the implicit interface in Origin suggests that some spatial references are made without explicit relation to another element. This is possibly due to the shared understanding or the context of the tasks, which revolve around the interface as assumed Origin.

Figure 5b shows the relations and frequencies of the themes of Target and Motive. Showing the motivation of participants during the study, the most frequent Motives were GUI information, map information, and POI information. This is correlating with the tasks, one being the GUI exploration and the other being a task

Table 1: All Themes and Categories of Spatial Referencing Utterances with examples from the study or further descriptions

Theme	Category	Utterance Examples or Descriptions	Occurrences Frequency
Target	Points of Interest	Specific locations, landmarks, or destinations of interest, "lake", "charging station"	38.02%
	GUI Element	GUI components or other interactive elements, "icon", "button", "widget"	38.02%
	Element Description	Non-specific descriptions about the appearance of visual content without necessarily naming the element, "area", "options"	26.44%
	Traffic and Navigation	Elements related to the navigation context, such as displayed routes or traffic conditions "route", "red traffic sections"	1.65%
Information	Absolute	Locations that can be anchored to the overall interface, "in the middle of the screen...", "on the display..."	44.63%
	Relative Direction	Direction described relative to another element; "under...", "south of...", "right of..."	26.45%
	Relative Point	Position described relative to another element; "at the bottom right of...", "right next..."	23.97%
	Icon Description	"burger icon", "arrow"	14.87%
	Size	"big", "smaller"	9.92%
	Interaction Context	References to the participant's interaction state; "the ones I'm seeing right now..."	9.10%
	Color	"blue", "green"	8.26%
	Textual Description	On- screen text used as identifier; "button with the N", "icon with 100", "the one that says SEARCH"	6.61%
	Function	Functionality associated with the element rather than their appearance, "the one you can search with", "ventilation symbol"	4.96%
	Shape	"elongated", "circles"	2.48%
	Name of GUI Element	Explicit name or label of the GUI element, "clock", "navigation menu"	2.48%
	GUI Touch Reference	Explicitly tied to involvement of touch interactions, "the point that I marked"	0.83%
Origin	Map	Reference anchored to the displayed map; "... of the map"	33.06%
	Interface	The overall display or screen; "...of the display", "...to the screen"	27.27%
	Points of Interest	Locations, landmarks or names of cities; "lake"	17.36%
	Implicit Interface	No explicit mentioned Origin, but the implicit interface as assumption	15.70%
	Text	On-screen text read and cited diractly; "button that says SEARCH"	12.40%
	Icon	"compass icon", "mountain icon"	5.79%
	1st Person Perspective	Participant's own viewpoint as anchor, "left of me..."	2.48%
Motive	GUI Information	Function or meaning of interface elements; "What does this button do..."	46.29%
	POI Information	Details about a POI; "What's the name of this lake..."	26.45%
	Map Information	Requests aimed at interpreting map content; "What's this blue lake on the map..."	16.53%
	Relocation of Map	Instructions to move or change the map in some form; "Show me on the map"	11.58%
	Navigation	Instructions to start a navigation or modify a route; "I want to get to xy", "Select the left route for me"	4.13%
	Correction	Revising or clarifying a prior utterance; "No, I meant...", "I was talking about..."	3.31%

aimed at asking for a POI. Results reveal a strong interconnectedness between GUI elements and their associated information, as well as POIs and their related context. There is a strong connection between a GUI element and GUI information. Example utterances were *"What's that icon upper right, next to the clock?"* or *"What can I do with the arrow next to search?"*. This indicates that asking about a GUI element often also aims to obtain information about its function or meaning. Similarly, POI references are strongly linked to inquiries about POI details such as names, ratings, or other contextual information, as well as to navigation and map manipulation. These connections of the GUI element and POI show that most of them are referenced with the motivation of staying in the same topic and to be maintaining a topical continuity. Map information is shown to be the most flexible, with a connection to all Target categories. Anything map-related is included in map information, like *"What is currently shown on the right side of the map?"* or *"What do the dark blue areas on the map mean?"*.

7 Discussion

The following section discusses the implications of our findings for the design and reliability of spatial referencing. We reflect on both technical implications and user interaction patterns that shape spatial referencing in automotive contexts.

7.1 Reliability of MLLMs for Spatial Grounding

Our evaluation of the prototype highlights both the potential and the current limitations of MLLMs, specifically using the GPT-4o model, when used for spatial grounding in automotive interfaces. While rule-based or existing systems with explicit access to the GUI state information could achieve higher technical reliability, our goal was not to evaluate system performance but to investigate user interaction patterns enabled by emerging MLLM capabilities. MLLMs were chosen because they can process both textual and visual inputs, allowing participants to use natural, unconstrained spatial references instead of predefined command structures, which might not be understood in a purely deterministic implementation.

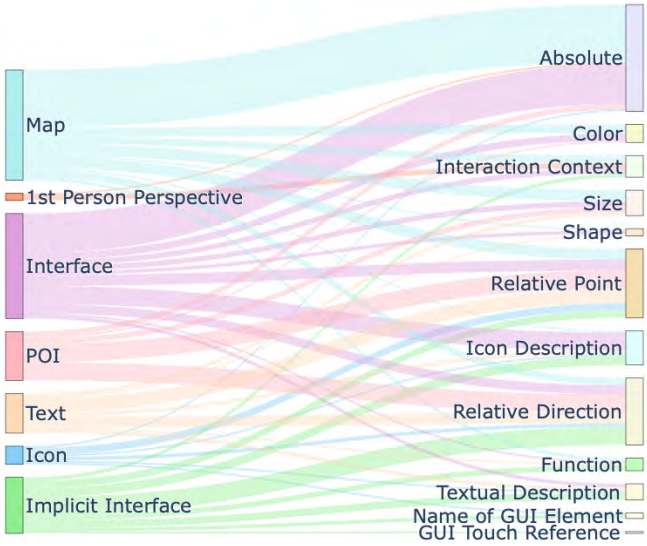


Figure 6: Frequencies and relationships of the two themes where Origin is referenced by Information.

Out of 116 spatial references, 72 responses (62.07%) were classified as correct, showing that the system is able to successfully interpret visual context and link user utterances to GUI or map elements in more than half of the cases. However, the remaining instances reveal several reliability issues that must be considered. A central challenge lies in the consistency of tool invocation. The model occasionally invoked the spatial grounding tool when it was not needed (23 instances) or failed to do so when appropriate (73 instances). Almost all these omissions occurred because a screenshot was already present in the context. While this mostly did not lead to incorrect responses in our relatively static scenarios, the proportion of correct responses could likely have been higher if the model had consistently invoked the tool for each spatial reference. In other cases, the model simply agreed with the user instead of grounding the response, as in the example: *"But at the top right, that's the current time showing on the screen right?"* - to which the model replied, *"You are right, at the top right of the screen you can see the current time!"*. This behavior undermines reliability, since display content may change, and the lack of grounding risks inconsistencies.

Such issues illustrate the current technical limitations of MLLMs for spatial reasoning. To increase reliability, future approaches should focus on three directions: (1) improving the detection of when spatial grounding is needed to reduce missed and unnecessary tool calls, (2) improve models by fine-tuning with automotive GUI-specific data, and (3) providing transparency to the user, for example by highlighting the element that the system grounded its response on, so that errors can be made transparent to the user to be corrected. Beyond improving technical reliability, these findings also provide guidance for the design of spatial referencing. Understanding where spatial reasoning may fail helps designers anticipate and work on how to make system reasoning more transparent.

7.2 Effects of Task Context on Spatial References

The proposed taxonomy demonstrates that spatial references are structured through four key themes: Target, Information, Origin, and Motive. Results show that 15.61% (116 out of 743) utterances contained spatial references during the study. Targets most frequently include Points of Interest (POIs) and GUI elements, with information conveyed through absolute and relative positioning. Maps and interfaces emerged as primary reference Origins. Motives centered on seeking GUI information, POI details, and map understanding. It is noteworthy that the nature of the tasks involved a heavy focus on the display. The two tasks in question are one with the task of asking about a specific POI displayed on the screen, and the other of exploring inside the GUI. Resulting in 20.07% (55 out of 274) spatial references in the second task and 20.63% (59 out of 286) spatial references during the third task. This demonstrates a task-dependent pattern in spatial reference usage, which correlates with the most frequent motives: GUI information and POI information. The first task, which was about finding a suitable charging station, only showed a quote of 1.09%, suggesting that participants did not require the use of references to the screen when searching for charging stations. This finding reveals a distinction between task types. POI search that can be accomplished through voice commands alone versus tasks that require visual exploration and reference to on-screen elements. Participants seemed to be satisfied with only the voiced information in combination with the textual information in the context of POI search. The only scenarios showed the need for *"show me the charging station on the map,"* so taking actions inside the map and relocating it.

7.3 User Strategies for Spatial Referencing and Topical Continuity

Insights from the results reveal the diversity of referencing strategies employed by participants. For instance, GUI elements were often referenced with the motive of understanding their function, as evident from utterances like "What's that icon upper right, next to the clock?" or "What can I do with the arrow next to search?." On the other hand, POIs were frequently linked to navigation and map interactions. The strong connection between Targets and Motives suggests that users maintain topical continuity through spatial references (see Figure 5b). For example, when referencing a GUI element, the motive was often to seek information about that element's functionality or purpose within the interface. Similarly, POI references were commonly associated with motives related to navigation, map manipulation, or gathering details about the specific POI. When tasked with referencing icons, participants often used other GUI elements, like other icons inside the GUI, as a reference. There seems to be a mental connection between the type that is referenced and the object used as a relation. When it comes to referencing strategies, the map, the interface, or the implicit interface as a whole was frequently used as the Origin to make spatial references, see Figure 6. This tendency suggests that participants preferred to refer to the interface holistically with directional cues like "left" or "right" instead of taking the additional mental step of describing the specific element in further detail. For instance, participants would say "on the map" or "on the display"

rather than providing a more granular description of the Target element's location or context within the interface. Our analysis also reveals users mostly used absolute positioning (44.63%) and relative directional (26.45%) or relative point (23.97%) references. This is often used in combination with the holistic approach mentioned before. The use of absolute positioning by phrases like "in the middle of the screen" or "on the display" suggests that users tend to reference within the overall spatial framework of the interface. Relative references were the second most common strategy, showing how users create relationships between elements to disambiguate their spatial references. In general, users opt for broader reference frames that require less cognitive effort to formulate, especially in context with VUIs, where there feels like there is less time to think.

7.4 From Taxonomy to Insights from Spatial Referencing Patterns

The taxonomy provides a structured way of analyzing how users formulate spatial references in in-car interactions. By decomposing utterances into Target, Information, Origin, and Motive, it becomes possible to identify recurring strategies and contextual dependencies for future design considerations. Further analysis with relationships and frequencies of spatial reference categories also implies the most associated themes. Beyond its descriptive value, the taxonomy also offers implications for in-car VUIs. First, the strong alignment between Target and Motive categories suggests that systems can anticipate likely user intents (e.g., GUI elements typically prompt questions about functionality, POIs about names or details). Second, the dominance of absolute and relative positioning indicates that systems should be optimized to interpret broader references such as "on the left" or "in the middle" rather than interpreting precise and descriptive formulations. Third, the frequent reliance on implicit frames of reference highlights the importance of treating the display or map as a default origin, reducing the need for users to explicitly specify context. In addition to informing design, the taxonomy can also support model performance. Its structured categories can provide a blueprint for constructing higher-quality datasets that better capture how users naturally reference in automotive contexts. While these implications are rooted in in-car interactions, the underlying mechanisms of spatial references are not domain-specific. Researchers developing spatial referencing for areas like robotics or augmented reality could adopt this taxonomy to train or evaluate models that interpret human spatial language under dynamic visual conditions. For example, synthetic utterances could be generated following the taxonomy's structure, or human-annotated datasets could be aligned with its categories to create utterance–response pairs. Such datasets would enable fine-tuning of MLLMs on automotive-specific GUIs and map-based applications.

7.5 Cognitive Demands in Voice-Based Interactions

During the study, participants had the opportunity to explore the interface. Even under such conditions, participants predominantly employed cognitively efficient strategies such as absolute references (e.g., "in the middle of the screen") or relative references (e.g., "to the right of the burger icon"). This suggests a preference for low-effort expressions in voice-based spatial referencing. Speech interaction itself can be seen as a constraint, as long breaks will cancel the input time window. Time pressure after pressing the push-to-talk button and the need for input likely reinforced the preference for short, absolute, or relative expressions over more elaborate descriptions of elements. Another factor may be participants' limited prior experience with automotive-specific voice assistants ($M = 3.05$) compared to automotive-specific assistants ($M = 1.38$). This could have reduced participants' confidence in making more complex and automotive-specific references. Although participants were introduced through an example task, they had little time during the study setup to develop strategies before starting the study. This can make it challenging to adapt to the learning curve and create a barrier of not knowing what is possible. As users become more comfortable with the new paradigm over time, they can adapt and interact with the feature more naturally. These findings show how cognitive constraints, such as time pressure or limited feedback, shape the linguistic strategies users employ when communicating spatial references. These insights can inform the design of voice- and vision-based interfaces beyond automotive use, where similar temporal and attentional demands exist.

7.6 Limitations

While enabling spatial referencing for LLMs presents opportunities for driver-vehicle interaction, several limitations must be acknowledged. The limitations of this work are resolved around model capabilities and the restricted scope of our current implementation.

First, the current output generation of the model in use (GPT-4o) shows inconsistencies, especially in spatial reasoning. Even when provided with a contextual image input, models are heavily reliant on the textual information, either taking the structure of the given object format in the data or plain text to generate output [41]. An example is "the route in the middle", where the reference is to a displayed route on the display in contrast to the textual route data object. Consequently, our study cannot conclude that the LLM genuinely grounds spatial references in visual content alone. The results reflect performance in a hybrid setup where textual cues remain dominant. Although our prompts already included examples and counterexamples of how spatial references should be resolved, the model still showed unreliable behavior in deciding when to invoke the tool. One contributing factor is that we built on top of an existing LLM backend that already provides a large set of tools. This suggests that limitations go beyond simple prompt engineering and relate to the model's underlying mechanism for tool invocation [35, 36]. More prompting strategies or fine-tuning, improved orchestration across available tools, or fine-tuning with domain-specific data to reduce such errors could reduce such errors.

Additionally, the integration was limited to the navigation application. Integrations of the touch modality and all other components that can be seen inside and outside of the vehicle (e.g., secondary displays, haptic buttons) were not included. The implementation is restricted to the display of the navigation system and what is shown there. The minimalistic design of the tested navigation GUI might not give enough content to reference spatially. By restricting spatial

referencing to the central display and within the navigation application, the system misses opportunities for more comprehensive and intuitive interactions across the entire vehicle interface.

7.7 Future Work

We conducted the study in a stationary vehicle providing a controlled environment to observe user behaviors. Future work should extend to dynamic driving conditions, where divided attention and changing map content create additional challenges. Spatial references such as "on the left" may quickly shift or disappear as the map updates. Here, transparency (e.g., a VUI-specific GUI that highlights referenced elements) could reduce cognitive load and allow users to verify or correct grounding, thus supporting trust in the system.

While our prototype focused on spatial references within the graphical interface, future research could explore how spatial language and multimodal input, such as gesture or gaze, can integrate across modalities and environmental contexts. This consideration becomes increasingly important as users may not clearly distinguish the system's perceptual boundaries. If their mental model assumes that the system "sees" the same visual layer as they do, spatial references may naturally extend beyond the in-car central information display.

Second, improvements in spatial reasoning are expected as research in MLLMs progresses. Fine-tuning with domain-specific GUI data could reduce hallucinations and improve grounding reliability beyond what prompting alone can achieve [25]. Recent work shows that models trained on image–text data benefit more from few-shot learning [6]. Interleaved image-text data also describes relationships within an image. Instead of image-text pairs like 'a giraffe," captions for interleaved image-text data would be a giraffe standing behind a metal fence. The giraffe appears to be looking towards the ground." Targeted pre-training and fine-tuning with automotive-specific examples could therefore strengthen reliability in handling spatial references.

Finally, spatial language is culturally and linguistically diverse. Even within English, dialects differ in how spatial relations are expressed [10]. Expanding to other languages and cultural contexts would provide valuable insights for designing adaptive systems that accommodate such variations.

8 Conclusion

In this work, we examined the integration of spatial referencing for LLM-based VUIs for automotive navigation. Through an in-vehicle study, we analyzed user interactions with visual content on navigation screens, highlighting how users refer to and describe on-screen elements with a taxonomy. The creation of a taxonomy for spatial references provides insights into the different ways users interact with visual information, with a preference for absolute and relative spatial references. Most users frequently rely on absolute and relative descriptions, often drawing on salient traits or nearby elements as anchors for their reference. While the study demonstrates an implementation of spatial referencing in automotive contexts, it also shows several challenges. Future research directions could focus on improving spatial reasoning capabilities, exploring cultural and linguistic differences in spatial descriptions, and investigating the challenges of spatial referencing in dynamic driving scenarios. As vehicles evolve towards higher levels of automation, the role of in-vehicle interfaces may shift, potentially increasing the importance of intuitive spatial referencing. Spatial referencing can play a central role by allowing users to seamlessly connect spoken language with visual and environmental context. The integration of spatial referencing in automotive interfaces provides the potential to make in-vehicle VUI and GUI interaction truly hands-free, more natural, and more human-like.

Author Contributions

Khanh Huynh: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Resources, Software, Visualization, Writing – original draft; **Jeremy Dillmann:** Conceptualization, Formal Analysis, Funding acquisition, Methodology, Supervision, Validation, Visualization, Writing – review & editing; **Sven Mayer:** Conceptualization, Formal Analysis, Methodology, Supervision, Validation, Writing – review & editing.

Acknowledgments

This work has been partly supported by the Research Center Trustworthy Data Science and Security (https://rc-trust.ai), one of the Research Alliance centers within the UA Ruhr (https://uaruhr.de).

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. IEEE, New York, NY, USA, 2425–2433. doi:10.1109/ICCV.2015.279

[2] Gilles Baechler, Srinivas Sunkara, Maria Wang, Fedir Zubach, Hassan Mansoor, Vincent Etter, Victor Cărbune, Jason Lin, Jindong Chen, and Abhanshu Sharma. 2024. ScreenAI: a vision-language model for UI and infographics understanding. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence* (Jeju, Korea) *(IJCAI '24)*. Article 339, 11 pages. doi:10.24963/ijcai.2024/339

[3] Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2021. Weakly supervised relative spatial reasoning for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE, New York, NY, USA, 1908–1918. doi:10.1109/ICCV48922.2021.00192

[4] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. *Qualitative Hci Research: Going Behind the Scenes*. Morgan & Claypool Publishers, San Rafael, CA, USA.

[5] Gary E Burnett and J Mark Porter. 2001. Ubiquitous computing within cars: designing controls for non-visual use. *International Journal of Human-Computer Studies* 55, 4 (2001), 521–531. doi:10.1006/ijhc.2001.0482

[6] Mustafa Dogan, Ilker Kesen, Iacer Calixto, Aykut Erdem, and Erkut Erdem. 2024. Evaluating Linguistic Capabilities of Multimodal LLMs in the Lens of Few-Shot Learning. *arXiv preprint arXiv:2407.12498* (2024). doi:10.48550/arXiv.2407.12498

[7] Abbe Don, Susan Brennan, Brenda Laurel, and Ben Shneiderman. 1992. Anthropomorphism: from Eliza to Terminator 2. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Monterey, California, USA) *(CHI '92)*. Association for Computing Machinery, New York, NY, USA, 67–70. doi:10.1145/142750.142760

[8] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467. doi:10.1080/10447318.2018.1456150

[9] Boyu Gou, Ruohan Wang, Boyuan Zheng, Yanan Xie, Cheng Chang, Yiheng Shu, Huan Sun, and Yu Su. 2024. Navigating the digital world as humans do: Universal visual grounding for gui agents. *arXiv preprint arXiv:2410.05243* (2024), 34. doi:10.48550/arXiv.2410.05243

[10] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 14375–14385. doi:10.1109/CVPR52733.2024.01363

[11] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. IEEE, 10867–10877. doi:10.1109/CVPR52729.2023.01046

[12] Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, et al. 2023. ChatGPT for shaping the future of dentistry: the potential of multi-modal large language model. *International Journal of Oral Science* 15, 1 (2023), 29. doi:10.1038/s41368-023-00239-y

[13] Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 9161–9175. doi:10.18653/v1/2023.emnlp-main.568

[14] Linus Kendall, Bidisha Chaudhuri, and Apoorva Bhalla. 2020. Understanding technology as situated practice: everyday use of voice user interfaces among diverse groups of users in urban India. *Information Systems Frontiers* 22 (2020), 585–605. doi:10.1007/s10796-020-10015-6

[15] Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications, Los Angeles, CA, USA. doi:10.4135/9781071878781

[16] Olga Lezhnina and Gábor Kismihók. 2020. A multi-method psychometric assessment of the affinity for technology interaction (ATI) scale. *Computers in Human Behavior Reports* 1 (2020), 100004. doi:10.1016/j.chbr.2020.100004

[17] Shen Li, Rosario Scalise, Henny Admoni, Stephanie Rosenthal, and Siddhartha S. Srinivasa. 2016. Spatial references and perspective in natural language instructions for collaborative manipulation. In *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'16)*. IEEE Press, New York, NY, USA, 44–51. doi:10.1109/ROMAN.2016.7745089

[18] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 292–305. doi:10.18653/v1/2023.emnlp-main.20

[19] Fangyu Liu, Guy Emerson, and Nigel Collier. 2023. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics* 11 (2023), 635–651. doi:10.1162/tacl_a_00566

[20] Amama Mahmood, Junxiang Wang, Bingsheng Yao, Dakuo Wang, and Chien-Ming Huang. 2025. User interaction patterns and breakdowns in conversing with llm-powered voice assistants. *International Journal of Human-Computer Studies* 195 (2025), 103406. doi:10.1016/j.ijhcs.2024.103406

[21] Manuel Masseno, Inês Lopes, Rita Marques, Francisco Rebelo, Elisângela Vilar, and Paulo Noriega. 2023. The Impact of Tangibility in the Input of the Secondary Car Controls: Touchscreens vs. Physical Buttons. In *International Conference on Design and Digital Communication*. Springer, Cham, Switzerland, 174–183. doi:10.1007/978-3-031-47281-7_14

[22] Christine Murad, Heloisa Candello, and Cosmin Munteanu. 2023. What's The Talk on VUI Guidelines? A Meta-Analysis of Guidelines for Voice User Interface Design. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (Eindhoven, Netherlands) *(CUI '23)*. Association for Computing Machinery, New York, NY, USA, Article 19, 16 pages. doi:10.1145/3571884.3597129

[23] Christine Murad, Cosmin Munteanu, Leigh Clark, and Benjamin R. Cowan. 2018. Design guidelines for hands-free speech interaction. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (Barcelona, Spain) *(MobileHCI '18)*. Association for Computing Machinery, New York, NY, USA, 269–276. doi:10.1145/3236112.3236149

[24] Mark B Neider, Xin Chen, Christopher A Dickinson, Susan E Brennan, and Gregory J Zelinsky. 2010. Coordinating spatial referencing using shared gaze. *Psychonomic bulletin & review* 17 (2010), 718–724. doi:10.3758/PBR.17.5.718

[25] OpenAI. 2024. Model optimization. https://platform.openai.com/docs/guides/model-optimization Accessed: August 08, 2025.

[26] Florian Roider, Sonja Rümelin, Bastian Pfleging, and Tom Gross. 2017. The Effects of Situational Demands on Gaze, Speech and Gesture Input in the Vehicle. In *Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications* (Oldenburg, Germany) *(AutomotiveUI '17)*. Association for Computing Machinery, New York, NY, USA, 94–102. doi:10.1145/3122986.3122999

[27] Albrecht Schmidt, Anind K. Dey, Andrew L. Kun, and Wolfgang Spiessl. 2010. Automotive user interfaces: human computer interaction in the car. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI EA '10)*. Association for Computing Machinery, New York, NY, USA, 3177–3180. doi:10.1145/1753846.1753949

[28] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems* (New Orleans, LA, USA) *(NIPS '22)*. Curran Associates Inc., Red Hook, NY, USA, Article 1833, 17 pages.

[29] Lukas Stappen, Jeremy Dillmann, Serena Striegel, Hans-Jörg Vögel, Nicolas Flores-Herr, and Björn W Schuller. 2023. Integrating generative artificial intelligence in intelligent vehicle systems. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, New York, NY, USA, 5790–5797. doi:10.1109/ITSC57777.2023.10422003

[30] Gareth Terry, Nikki Hayfield, Victoria Clarke, Virginia Braun, et al. 2017. Thematic analysis. *The SAGE handbook of qualitative research in psychology* 2, 17-37 (2017), 25.

[31] Gerhard van Huyssteen, Aditi Sharma Grover, and Karen Calteaux. 2012. *Voice user interface design for emerging multilingual markets*. Sun Press, London, UK, 291.

[32] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. 2025. Is A Picture Worth A Thousand Words? Delving Into Spatial Reasoning for Vision Language Models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS '24)*. Curran Associates Inc., Red Hook, NY, USA, Article 2400, 30 pages.

[33] Maximiliane Windl, Verena Winterhalter, Albrecht Schmidt, and Sven Mayer. 2023. Understanding and Mitigating Technology-Facilitated Privacy Violations in the Physical World. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 585, 16 pages. doi:10.1145/3544548.3580909

[34] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, New York, NY, USA, 2247–2256. doi:10.1109/BigData59044.2023.10386743

[35] Hongshen Xu, Zichen Zhu, Lei Pan, Zihan Wang, Su Zhu, Da Ma, Ruisheng Cao, Lu Chen, and Kai Yu. 2024. Reducing tool hallucination via reliability alignment. *arXiv preprint arXiv:2412.04141* (2024). doi:10.48550/arXiv.2412.04141

[36] Seungbin Yang, ChaeHun Park, Taehee Kim, and Jaegul Choo. 2024. Can Tool-augmented Large Language Models be Aware of Incomplete Conditions? *arXiv preprint arXiv:2406.12307* (2024). doi:10.48550/arXiv.2406.12307

[37] Yuhao Yang, Yue Wang, Dongxu Li, Ziyang Luo, Bei Chen, Chao Huang, and Junnan Li. 2024. Aria-ui: Visual grounding for gui instructions. *arXiv preprint arXiv:2412.16256* (2024). doi:10.48550/arXiv.2412.16256

[38] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* (2024), nwae403. doi:10.1093/nsr/nwae403

[39] Keen You, Haotian Zhang, Eldon Schoop, Floris Weers, Amanda Swearngin, Jeffrey Nichols, Yinfei Yang, and Zhe Gan. 2024. Ferret-UI: Grounded Mobile UI Understanding with Multimodal LLMs. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXIV*. Springer-Verlag, Berlin, Heidelberg, 240–255. doi:10.1007/978-3-031-73039-9_14

[40] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it?. In *The Eleventh International Conference on Learning Representations*. 20. doi:10.48550/arXiv.2210.01936

[41] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?. In *European Conference on Computer Vision*. Springer, Cham, Switzerland, 169–186. doi:10.1007/978-3-031-73242-3_10