A Comparative Study of Human-operated and AI-driven Guidance with a Teleoperated Mobile Robot

Ao Guo[1], Shota Mochizuki[1], Sanae Yamashita[1], Kenya Hoshimure[2], Jun Baba[2], Ryuichiro Higashinaka[1]
[1]Graduate School of Informatics, Nagoya University, Japan
[2]AI Lab, CyberAgent, Japan
Correspondence: guo.ao.i6@f.mail.nagoya-u.ac.jp

Abstract

Recent advances in large language models (LLMs) such as GPT-4o offer the potential for enhancing AI-driven robotic interactions, but their effectiveness in mobile tour guidance remains unexplored. This study investigates the differences between human-operated and AI-driven guidance at an aquarium using Teleco, a teleoperated mobile robot, in a real-world field experiment. A total of 277 guidance sessions were collected under two modes: human-operated, where the operator controlled all dialogue, actions, and movement, and AI-driven, where GPT-4o generated responses while the operator only controlled the robot's actions and movement. Our results indicate that human-operated guidance places greater emphasis on visitor movement, spatial positioning during observation guidance, and empathetic expressions, whereas AI-driven guidance promotes conversational engagement by frequently prompting visitors to ask questions. In addition, we found that user behaviors, including users' gaze patterns and vocabulary richness, also serve as valuable indicators reflecting their overall experience during guidance interactions. Furthermore, empathetic expression is recognized as the key differentiating factor between the two guidance modes, significantly influencing users' overall experience.

(a) Operator on 3rd floor teleoperating Teleco

(b) Teleco interacting with visitors on 1st floor

Figure 1: Photographs of Teleco providing mobile guidance under teleoperation by a remote operator.

1 Introduction

With advances in robotics technology, robots are increasingly being deployed in various public spaces, such as airports for information services (Triebel et al., 2016) and museums or aquariums for visitor guidance (Mochizuki et al., 2023; Yamashita et al., 2023b). Recent advancements in situation-aware technologies have further contributed to the growing adoption of mobile robots, as they offer greater adaptability and interactivity compared to stationary robots. In venues such as museums and aquariums, mobile robots have the potential to provide more complex tour guidance by dynamically responding to visitors and their surroundings.

However, despite extensive research on autonomous navigation and context-aware interactions (Chen et al., 2021; Han and Li, 2023; Al-Kamil and Szabolcsi, 2024), mobile tour guide robots still face major limitations in their dialogue ability. Many existing robots use pre-defined templates for dialogue generation (Yuguchi et al., 2022; Vásquez and Matía, 2020), which limits flexibility and reduces the naturalness of the interactions. The emergence of large language models (LLMs) (Achiam et al., 2023) has demonstrated promising advancements in natural language understanding and conversational AI, making them a potential solution for improving robotic guidance. However, few studies have explored integrating LLMs into mobile guidance robots. This raises two research questions:

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 1615–1627

RQ1: To what extent can high-performing LLMs (e.g., OpenAI's ChatGPT (Roumeliotis and Tselikas, 2023)) provide guidance using a teleoperated mobile robot compared with human-operated performance?

RQ2: What are the differences between human-operated and AI-driven mobile robot guidance regarding user experience and interaction quality?

To investigate the differences between human-operated and AI-driven guidance, we conducted a field experiment using Teleco[1], a teleoperated mobile robot, at Nifrel[2], a facility that combines an aquarium and a zoo in Osaka, Japan. Teleco provided a 5-minute tour, introducing three pufferfish species across two exhibit tanks (see Fig. 1).

The experiment involved two guidance modes: one where a human operator directly controlled the robot's dialogue, actions, and movement, and another where an AI-driven system implemented by GPT-4o generated responses while the operator only controlled the robot's actions and movement. Both modes were conducted under a Wizard-of-Oz setup, in which the human operator controlled the non-autonomous functions corresponding to each mode. During the experiment, participants were not informed of the operator's involvement in either mode, nor of which parts were automated or manually controlled. To assess the capability of AI-driven guidance, we evaluated post-guidance questionnaires. To further distinguish the differences between the two modes, we analyzed interaction patterns using n-gram frequency analysis, identified multimodal features correlated with questionnaire responses through correlation analysis, compared utterance categories via embedding-based clustering, and examined relationships influencing user experiences through structural equation modeling (SEM).

2 Related Work

Mobile robots are increasingly utilized for human interaction in various guidance tasks, such as visitor assistance in museums and aquariums, offering a flexible guidance experience. One of the earliest robots for museum guidance, *MINERVA* (Thrun et al., 2000), was designed for human-machine interaction. Similarly, *Robovie* (Ishiguro et al., 2001) featured human-like actuators, vision, and audio sensors, enabling it to exhibit human-like behaviors. Honda's *ASIMO* (Nakano et al., 2005) further advanced mobility, enabling it to walk and to perform receptionist and information guide tasks.

More recently, the *Pepper* robot (Pandey et al., 2018) was designed with emotion expression capabilities and wheeled mobility, making it well-suited for customer interactions in service environments. *Doris* (Vásquez and Matía, 2020) was developed as a tour guide robot capable of autonomously navigating predefined waypoints while displaying emotional expressions. Its guidance relied on several predefined templates for dialogue generation. Similarly, *Butsukusa* (Yuguchi et al., 2022) integrated environment recognition and object detection of surroundings to enable autonomous movement in indoor spaces. However, its dialogue generation remained template-based.

With the emergence of LLMs, researchers have explored embedding them into mobile robots (Zeng et al., 2023). Shah et al. (2023) integrated a pre-trained LLM (GPT-3) to process language, vision, and actions for navigation. Similarly, Liu et al. (2024) developed a robot with ChatGPT that can perceive its surroundings and perform tasks while communicating with humans. *PaLM-SayCan* (Brohan et al., 2023) grounds natural language instructions into executable actions by combining a language model with affordance functions. *PaLM-E* (Driess et al., 2023) represents one of the first embodied multimodal LLMs, unifying vision, language, and state inputs to generate both textual and action outputs. *RT-1* (Brohan et al., 2022) leverages vision-language LLMs by encoding visual and language inputs into tokens and training a transformer to perform over 700 real-world tasks, demonstrating strong multi-task generalization.

While these studies have made advancements with robot perception, planning, and control, little attention has been paid to user experience and interaction quality. In the current study, we focus on the user experience and interaction quality by conducting a large-scale field experiment that directly compares human-operated and LLM-driven robot guidance.

3 Robot Systems for Human-operated and AI-driven Guidance

To compare AI-driven and human-operated guidance, we developed a teleoperated mobile robot

[1] https://www.vstone.co.jp/english/
[2] https://www.nifrel.jp/en/

system with two modes: human-operated and AI-driven. Teleco, a mobile robot equipped with an OLED display and wheeled mobility, was chosen to provide visitor guidance. The target language of the system was Japanese.

3.1 Human-operated Guidance

In this mode, the operator directly speaks to the visitor while simultaneously controlling Teleco's movements, expressions, gestures, and template utterances as needed. To ensure a seamless auditory experience, retrieval-based voice conversion (RVC)[3], a voice conversion model, is used to modulate the operator's voice and match it with Teleco's robotic tone, minimizing discrepancies when switching between direct speech and template utterances. To facilitate real-time control and smooth interactions, the teleoperation interface (Fig. 2) provides access to a live camera feed, dialogue history, visitor gaze data, recognized objects, robot gesture and expression controls, predefined utterance templates, and a navigation map.

3.2 AI-driven Guidance

In this mode, all utterances are generated by AI using GPT-4o (Achiam et al., 2023) based on automatic speech recognition (ASR) results via the Google Speech-to-Text API. A predefined prompt for GPT-4o is used (see Table 6 in Appendix A), which includes necessary information about the introduced content and a one-sentence visitor gaze description. This gaze description is automatically generated from visitor gaze data captured using an Insta360 GO 3S[5] mounted on a head-worn accessory. Note that we collected the user's head-mounted camera stream here, which was then utilized as an approximation of what the user was looking at rather than actual eye gaze. The captured gaze data is processed in real-time using GPT-4o-mini, a lighter variant of GPT-4o, to create concise gaze summaries while minimizing computational costs. These gaze descriptions are used for AI-driven dialogue generation, as user gaze plays a crucial role in guided tours, influencing both visitor engagement and information perception (Ruhland et al., 2015; Schreiter et al., 2023).

The AI-generated utterances are then converted into Teleco's robotic tone utilizing an embedded text-to-speech (TTS) system before being spoken aloud. The operator does not actively chat with the visitor in this mode; instead, their role is to monitor the conversation and manually control Teleco's movements, as it is difficult to navigate in congested environments. If the AI-generated dialogue contains incorrect information, the operator can pause the AI interaction and manually intervene to provide corrections.

Figure 2: Teleoperation interface. The interface includes (1) a live video feed from Teleco's chest camera, enabling the operator to monitor the surroundings; (2) a dialogue history transcribed via Google Speech-to-Text API[4]; (3) user gaze images streamed from the Insta360 GO 3S mounted on the visitor's head, providing information about visitor attention; (4) automatically recognized gaze objects with one-sentence descriptions extracted from user gaze images; (5) controls for Teleco's actions, including bowing, hand-raising, and height adjustments; (6) controls for five facial expressions (sad, happy, angry, nervous, and neutral); (7) controls for predefined template utterances, allowing the operator to quickly select consistent speech responses; and (8) movement controls via a map interface, allowing the operator to navigate the robot.

4 Field Experiment

To assess the capability of AI-driven guidance and clarify its differences from human-operated guid-

[3] https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI/

[5] https://www.insta360.com/product/insta360-go3s/

[4] https://cloud.google.com/speech-to-text/

ID	Questionnaire Items
Q1	I was satisfied with the conversation.
Q2	I actively participated in the conversation.
Q3	My interest and curiosity toward the introduced creatures deepened through the conversation.
Q4	I felt a sense of closeness with the robot.
Q5	The robot's speech was informative.
Q6	The robot's guidance was easy to understand.
Q7	The robot maintained an appropriate sense of distance.
Q8	The robot was looking at the same things as me.
Q9	The robot's speech and actions were consistent.
Q10	The robot's speech was appropriate for the situation.
Q11	The robot's actions were appropriate for the situation.

Table 1: User questionnaire items (translated from Japanese).

ance, we conducted a field experiment using the Teleco robot at Nifrel.

The experiment was conducted over 28 days. During this period, Teleco was controlled under two different modes, human-operated and AI-driven, to provide short guidance tours to visitors. Multimodal data, including audio dialogues, video recordings, and user gaze tracking data, were collected for further analysis. The field experiment was approved by our institution's ethics committee and conducted in compliance with ethical guidelines.

4.1 Experimental Workflow

Our study focused on guiding visitors through two exhibit tanks featuring three species of pufferfish. The experiment involved a single operator per session, with one male and one female operator assigned on different days. Both operators had native proficiency in Japanese and were trained in basic teleoperation as well as in the pufferfish species introduced during the tour. Staff members invited visitors to participate or responded to those who showed interest in the guidance tour.

Once the tour session began, Teleco moved sequentially from the small tank (containing Okinawa pufferfish) to the large tank (housing mouse pufferfish and makeup pufferfish), providing an interactive guidance experience. The robot introduced the three species and answered visitors' questions. At the end of the tour, Teleco returned to its initial position, and both the visitor and operator completed a post-guidance questionnaire to evaluate their guidance experience. The user questionnaire, consisting of 11 items, assessed user satisfaction, engagement, interest level, perceived closeness, and various aspects of the robot's performance (see Table 1). Each item was rated on a 7-point Likert scale, where 1 represented the lowest rating (worst) and 7 represented the highest (best). The operator questionnaire assessed the effectiveness of the guidance and the usability of the teleoperation system (see Table 7 in Appendix B). Each session lasted approximately five minutes, with flexibility to shorten or extend the duration based on visitor engagement and contextual factors such as congestion or ongoing events in the facility.

	Human	AI
No. of conversations	212	65
Total conv. time (h)	21.9	6.1
Avg. conv. time (sec)	372±101	336±112
Avg. no. of utterances	66.3±21.2	55.4±19.3
Avg. sys utt. duration (sec)	4.5±3.8	8.8±12
Avg. usr utt. duration (sec)	4.6±0.87	4.8±1.28

Table 2: Statistics of conversations under human and AI modes.

4.2 Statistical Information

We collected over 300 dialogues in this field experiment. After filtering out incomplete interactions caused by WiFi issues in congested conditions, a total of 277 dialogue sessions were retained, consisting of 212 human-operated sessions and 65 AI-driven sessions. The number of AI-driven sessions was lower due to environmental constraints; since speech recognition in the AI-driven mode required a quieter environment, these sessions were conducted only when background noise levels were low.

Table 2 presents a comparison of the dialogue statistics. All guidance utterances were transcribed using OpenAI Whisper-large-v3[6]. The results show that human-operated sessions had a longer average conversation time and more utterances than AI-driven sessions, indicating richer dialogue and interaction in the human-operated mode. In addition, the average system utterance duration in the human-operated mode was shorter than in the AI-driven mode, suggesting that AI-generated responses were longer and took more time to deliver.

5 Data Analysis

In this section, we analyze the differences between human-operated and AI-driven guidance from mul-

[6]https://huggingface.co/openai/whisper-large-v3

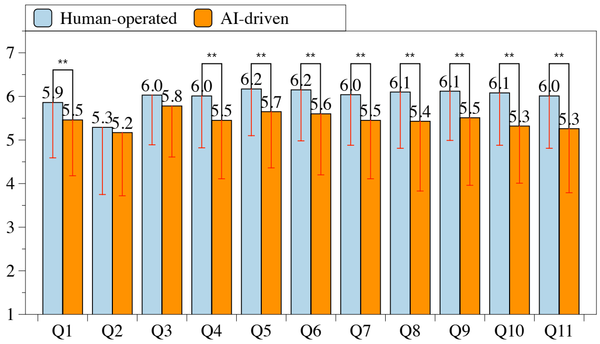

Figure 3: Results of user questionnaire, with each item rated on a Likert scale from 1 to 7.

tiple perspectives. First, we examine user questionnaire results to evaluate the overall performance of the two guidance modes. Next, we extract and compare lexical differences in system utterances. We then analyze correlations between questionnaire results and multimodal features derived from dialogue, user gaze behavior, and robot actions. Subsequently, robot utterances are categorized through embedding-based clustering to further identify key linguistic differences between the two modes. Lastly, we construct a four-layer SEM structure to examine the hypothesized directional relationships among these multimodal variables and their influences on user experiences.

5.1 Results of Post-guidance User Questionnaire

The post-guidance questionnaire results for human-operated and AI-driven guidance are shown in Fig. 3. On average, human-operated guidance received a score of 6.0, while AI-driven guidance scored 5.4. Across all questionnaire items, human-operated guidance consistently received higher ratings than AI-driven guidance. Since this study primarily focuses on user experience, operator questionnaire results are not analyzed here and are left for future investigation.

To statistically analyze the differences, we used a Mann-Whitney U test on each questionnaire item. The results showed significant differences between the two modes for all items except Q2 (user engagement) and Q3 (interest level). This suggests that AI-driven guidance showed the ability to achieve a human-level performance to some extent, although there is still a noticeable gap. The lack of significant differences in Q2 and Q3 indicates that both modes were equally effective in maintaining user engagement and sparking interest, possibly because the guidance content itself (three species of pufferfish) was engaging, and both modes provided essential information about the guidance content.

5.2 Lexical Differences in System Utterances

To clarify the specific linguistic differences in the dialogue between human-operated and AI-driven guidance, we performed an n-gram frequency analysis, following the approach of Yamashita et al. (2023a). We utilized word 4-grams in this analysis, as they provide a balance between capturing meaningful word sequences and maintaining interpretability, and analyzed the most distinctive 4-grams that exhibited significant differences between the two modes.

First, we utilized MeCab (Kudo et al., 2004), a morphological analyzer, to tokenize utterances and extract words from both human-operated and AI-driven dialogues. We then extracted 4-grams from the tokenized text. For each 4-gram, we compared its occurrences in human-operated and AI-driven dialogues by computing the proportion of its frequency relative to the total occurrences of all 4-grams in each mode. To determine whether the difference in relative frequencies was statistically significant, we applied the two-proportion z-test (Zou et al., 2003), which is suitable for comparing proportions between two independent samples.

The results of the distinctive 4-grams we obtained (see Table 8 in Appendix C) indicate that in the human-operated mode, Teleco frequently used directive phrases such as "Follow me" and "Please come here", emphasizing spatial guidance. In contrast, in the AI-driven mode, Teleco tended to use more question-oriented statements such as "Feel free to ask" and "Ask me anything", highlighting a stronger focus on interactive engagement with the visitor. This pattern is also evident in the example dialogues for both modes provided in Table 9 in Appendix D. These examples suggest that human-operated guidance places more emphasis on visitor movement and spatial positioning, whereas AI-driven guidance, due to its limited ability to manage spatial guidance, had no alternative but to rely more on conversational engagement by prompting visitors to ask questions.

5.3 Correlation Analysis of Multimodal Features with User Questionnaire Results

To clarify the factors influencing users' questionnaire ratings, we analyzed the correlation between

Human-operated											
Multimodal Features	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**	**Q11**
Avg. sentence len (Sys)	.14	.10	.07	.06	.12	.11	.11	.10	.12	.18	.01
Lexical diversity (Sys)	−.05	−.03	−.02	−.01	.02	−.03	.04	.04	.01	−.06	−.05
Vocabulary (Sys)	.17	**.19**$^+$	.14	.08	.12	.11	.02	−.03	.03	.16	.15
Avg. sentence len (Usr)	.12	.12	.20	.06	.16	**.23***	.00	.12	.12	.12	.08
Lexical diversity (Usr)	.04	−.14	−.09	−.05	.04	−.07	−.04	−.06	−.06	−.05	−.07
Vocabulary (Usr)	.06	**.20**$^+$	.14	.10	.08	.16	.10	.11	.06	.08	.10
User gaze entropy	.10	.05	−.03	.10	.08	.03	.00	.07	.08	.10	.12
Total action count	−.05	−.01	−.01	.03	−.03	−.05	−.03	−.09	−.03	.01	.03

AI-driven											
Multimodal Features	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**	**Q6**	**Q7**	**Q8**	**Q9**	**Q10**	**Q11**
Avg. sentence len (Sys)	.10	.09	.25	.16	.17	.16	.18	.19	.12	−.09	.02
Lexical diversity (Sys)	−.34	−**.33**$^+$	−.20	−.12	−.06	−.05	−.09	−.13	−.13	.05	.03
Vocabulary (Sys)	.22	**.38**$^+$	.21	.08	.07	.06	.11	−.01	.10	−.08	−.05
Avg. sentence len (Usr)	−.01	−.16	−.08	−.12	−.12	−.09	−.27	.05	−.02	−.13	−.08
Lexical diversity (Usr)	−.05	−.25	.08	.07	.05	.08	.08	.18	.09	.09	.03
Vocabulary (Usr)	.16	.10	−.01	−.04	−.02	−.14	−.16	−.03	−.05	−.21	−.20
User gaze entropy	.27	.02	−.05	**.31***	−.02	−.05	−.06	.22	.05	.14	−.02
Total action count	.31	.08	.14	.12	.24	.25	.31	.34	.33	**.37**$^+$	.27

Refer to Table 1 for descriptions of Q1 to Q11. $+p < .1$, $*p < .05$

Table 3: Correlation results for human-operated and AI-driven guidance. Bold values indicate significant correlations after two-stage false discovery rate correction.

multimodal features extracted from dialogue, user gaze, and robot actions with user questionnaire results for both human-operated and AI-driven guidance.

Following previous work by Guo et al. (2024), for each guidance session, we extracted linguistic features from both system and user utterances, including average sentence length, lexical diversity, and vocabulary as key linguistic characteristics. We also calculated user gaze transition entropy (Krejtz et al., 2015), which quantifies the variability and unpredictability of gaze shifts between different regions of interest, providing information about attention distribution. In addition, we calculated the total action count from Teleco's action data, which includes the operator's control of its gestures, movements, and facial expressions. The total action count was utilized to evaluate the richness of the system's interactive behaviors.

Table 3 presents the Spearman's rank correlation coefficient (Sedgwick, 2014) results between user questionnaire items and multimodal features for human-operated and AI-driven guidance. A two-stage false discovery rate correction ($\alpha < 0.05$) was applied separately for each questionnaire item to evaluate the statistical significance of correlations. Note that for human-operated guidance, system linguistic features refer to those extracted from the operator's utterances, whereas for AI-driven guidance, they refer to features extracted from AI-generated utterances. Our key findings are as follows.

Language usage correlates with user engagement and guidance clarity. For Q2 (user engagement), both human-operated and AI-driven guidance were influenced by the richness of Teleco's vocabulary (the operator's utterance vocabulary in human-operated mode and the AI-generated utterance vocabulary in AI-driven mode), suggesting that in both modes, diverse language contributes to user engagement. Significant correlations for Q2 (user engagement) and Q6 (guidance clarity) indicated that, in human-operated guidance, longer user sentence length and richer vocabulary were associated with higher user ratings on these two dimensions.

User gaze and robot actions correlate with users' perceived familiarity with the robot and its speech appropriateness. In AI-driven guidance, ratings for Q4 (familiarity with robot) and Q10 (speech appropriateness) correlated with user gaze entropy and total action count, respectively. This suggests that users who actively explored their surroundings tended to feel more familiar with the robot, with the richness of robot actions further influencing their experience—likely due to the AI system's limited ability to adapt its guidance based on visitor movements. In contrast, human-operated guidance seems to have adapted to visitor behaviors, making dialogue the primary factor influencing user experience.

To better understand the association between robot actions and Q10 in AI-driven guidance, we compared the total action count using a Mann–Whitney U test. The human-operated mode showed significantly higher counts (median = 6.0) than the AI-driven mode (median = 4.0; $p < .01$). In the human-operated condition, frequent use of actions (facial expressions and gestures) was common and may therefore have been less influential on Q10. In contrast, in the AI-driven condition, actions were fewer. We conclude that operators may have needed to focus on listening to the AI-generated dialogue and on navigation. Accordingly, each action may have been more salient and therefore more closely associated with Q10.

5.4 Utterance Clustering Using Sentence Embeddings

To explore differences between human-operated and AI-driven guidance regarding typical utterance categories, we conducted clustering analyses on system utterances extracted from transcribed dialogue data. Each utterance from both guidance modes was first embedded using the Japanese Sentence-BERT model[7]. We then applied K-means clustering with Euclidean distance to these Sentence-BERT embeddings. The optimal number of clusters was selected from a range of 2 to 10, determined by the silhouette score (Shahapure and Nicholas, 2020). For each cluster, we selected five utterances closest to the centroid as representative examples, provided them to GPT-4o, and instructed it to generate a concise description for each cluster.

The results of utterance clustering and their descriptions are summarized in Table 4. Operator utterances were clustered into nine categories, and a cluster proportion analysis was conducted to examine differences between human-operated and AI-driven guidance across these nine utterance categories. Specifically, we calculated the proportion of utterances in each category separately for human-operated and AI-driven guidance, then compared these proportions using two-proportion z-tests. The Benjamini–Hochberg (FDR-BH) procedure with $\alpha = 0.05$ was applied to control the false discovery rate across multiple comparisons.

The results in Table 5 reveal that, compared to AI-driven guidance, human-operated guidance provided more detailed introductions of the Okinawa pufferfish (the first species introduced), expressed greater empathy, offered more observation guidance, and used fewer movement instructions and prompting questions. Given the generally lower ratings observed for AI-driven guidance, these findings suggest that users' overall experiences could be improved if dialogues contain more empathetic expressions, detailed observation guidance, and richer information about the initially introduced species. The inclusion of such richer information likely reflects and promotes greater user engagement from the beginning of guidance.

Compared with the results in Subsection 5.2, which indicate that human-operated guidance places greater emphasis on visitor movement and spatial positioning, Table 5 shows that AI-driven guidance contains a higher proportion of move in-

Utterance Categories and Examples	
Okinawa Pufferfish Introduction	That one is called an "Okinawa puffer", but it's actually a relative of the porcupinefish.
Toxicity Warning	Its entire body is poisonous, so it isn't edible.
Mouse Pufferfish Introduction	The black-spotted porcupinefish is a large pufferfish with dark mottled patterns along its body.
Move Instruction	All right, let's move on to the next area.
Empathy Expression	Lucky you! You got to see it all puffed up!
Behavior Induction	Certainly. I'll guide you. Please follow right behind me.
Makeup Pufferfish Introduction	As its name suggests, the make-up pufferfish has markings around its eyes that look like cosmetics.
Observation Guide	In this large tank, a humphead wrasse is swimming. Can you find it?
Prompting Question	Any guesses? I wonder if you can tell.

Table 4: Utterance categories generated by GPT-4o and representative examples (translated from the original Japanese).

Proportion of Utterance Categories			
Category	**Human**	**AI**	**Sig.**
Okinawa Pufferfish Intro.	**16.8%**	11.3%	**
Toxicity Warning	4.3%	4.3%	
Mouse Pufferfish Intro.	16.2%	**18.9%**	**
Move Instruction	9.6%	**11.9%**	**
Empathy Expression	**11.9%**	8.8%	**
Behavior Induction	13.6%	13.6%	
Makeup Pufferfish Intro.	**9.7%**	8.1%	*
Observation Guide	**10.3%**	7.3%	**
Prompting Question	7.4%	**15.9%**	**

* $p < .05$, ** $p < .01$ (FDR-BH corrected)

Table 5: Proportions of utterance categories in human-operated and AI-driven guidance. Bold values indicate the higher proportion for each category between the two groups.

[7] https://huggingface.co/sonoisa/sentence-bert-base-ja-mean-tokens-v2

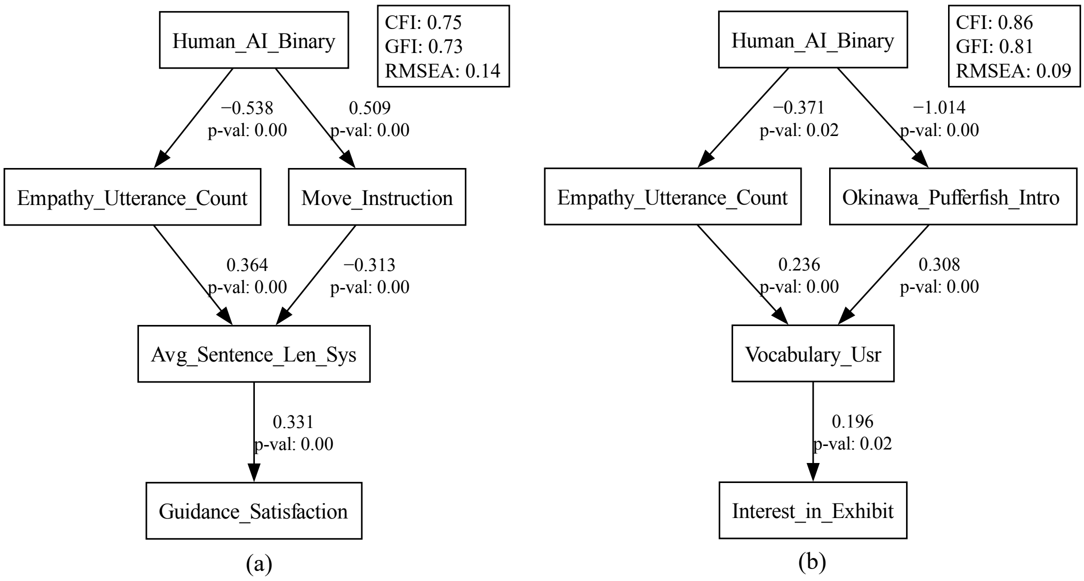

Figure 4: SEM results for two representative structures. The binary variable Human_AI_Binary is coded as 0 for human-operated guidance and 1 for AI-driven guidance.

structions. The distinction is that the AI tended to repeat the same transition phrase when moving to the next tank (e.g., "Let's move to the next area"), particularly in our field experiment at a public aquarium, where moving from the small tank to the big tank often required extra time in crowded conditions. In contrast, human operators used such instructions less frequently but produced more diverse spatial expressions throughout the guidance (e.g., "Follow me," "Please look here"). Consequently, the AI-driven system generated more move instructions in raw counts (11.9% vs. 9.6%) in the utterance proportion analysis, whereas the human operators provided richer and more varied spatial guidance, which made the move instructions more salient in the lexical analysis. Thus, the two analyses are complementary rather than contradictory.

5.5 SEM Analysis

Due to its capability to simultaneously analyze hypothesized directional relationships among multiple variables (Fan et al., 2016), we utilized structural equation modeling (SEM) to clarify differences between human-operated and AI-driven guidance regarding how multimodal behaviors influence user ratings. Referring to the previous study by Guo et al. (2025), which assumed a three-layer structure to clarify the influence of personality on task performance, we hypothesized a four-layer SEM structure with hierarchical relationships from top to bottom, where the first layer represents the guidance mode as a binary variable (0 for human-operated, 1 for AI-driven), the second and third layers correspond to utterance categories and multimodal features analyzed previously in Subsections 5.3 and 5.4, and the fourth layer is a specific questionnaire item targeted for analysis. The ordering of the second and third layers is based on the assumption that broader communication strategies (layer 2) guide the expression of specific multimodal behaviors (layer 3), which in turn influence users' overall experiences, such as satisfaction.

We applied regularized SEM (Jacobucci et al., 2016) to optimize the SEM structure. Specifically, we first implemented Lasso regression to select significant high-level and low-level features as predictors. Then, we built initial SEM models and iteratively applied a stepwise procedure, removing non-significant paths at each step until all remaining paths were statistically significant.

An individual SEM model was built for each of the 11 questionnaire items. Two typical questionnaire items, guidance satisfaction (Q3) and interest in the exhibit (Q6), were selected as representative examples and are shown in Fig. 4; the remaining questionnaire items exhibited similar patterns. In the first structure (left side of Fig. 4), compared with human-operated guidance, AI-driven guidance exhibited fewer empathetic expressions and provided more movement instructions, resulting in shorter average sentence lengths within dialogues and thereby decreasing guidance satisfaction. In the second structure (right side of Fig. 4), AI-driven guidance again showed fewer empathetic expressions and provided less introduction about the initially presented species (Okinawa pufferfish), leading to poorer user vocabulary and consequently reducing users' interest in the exhibit. These findings highlight empathetic expression as the key difference between the two guidance modes, consistent with the results presented in Table 5, where human-operated guidance demonstrated greater empathy and received higher overall user ratings.

6 Conclusion and Future Work

In this study, we conducted a comparative analysis of human-operated and AI-driven guidance using Teleco, a teleoperated mobile robot. Through a field experiment, we collected and analyzed multimodal data including linguistic features, user gaze behavior, and robot actions to investigate the differences between human and AI-driven interactions.

We evaluated the results of a user questionnaire to address RQ1 (How effective is AI-driven guidance compared to human-operated performance?) and concluded that while AI-driven guidance does not yet fully match human-operated performance, it achieves a comparable level in certain aspects. Moreover, our analysis addressing RQ2 (What are the differences between human-operated and AI-driven guidance?) revealed that human-operated guidance places more emphasis on visitor movement, spatial positioning during observation guidance, and empathetic expressions, whereas AI-driven guidance promotes conversational engagement by frequently prompting visitors to ask questions. In addition, user behaviors, including users' gaze patterns and vocabulary richness, can serve as valuable indicators reflecting their overall experience during guidance interactions. Furthermore, empathetic expression is recognized as the key differentiating factor between the two guidance modes, significantly influencing users' overall experience.

Several aspects of this study should be improved in future work. First, additional dialogue-level analyses, such as dialogue structure analysis, would further enrich the computational linguistics perspective of this work; multimodality, such as gaze, and its relationship to spatial movement cues may also help explain differences between human and LLM operators. Second, due to the stricter environmental conditions (i.e., lower background noise levels) required for accurate speech recognition in the AI-driven guidance mode, fewer AI-driven guidance sessions were collected compared to human-operated sessions. Future studies should include more AI-driven sessions to enable more reliable comparisons. Third, since the robot's movement was restricted to predefined waypoints, a more flexible navigation system should be implemented to enhance adaptability and interaction quality. Fourth, since the human-operated condition also contains a subset of sessions collected in quiet environments comparable to the AI-driven mode, it is possible to perform a controlled comparison under similar noise levels. We plan to conduct a more systematic comparison with this subset in future work. Fifth, in the AI-driven mode, the perception design mainly relied on gaze input, which constrained the system's ability to generate more contextually grounded and empathetic utterances. In future work, we plan to extend the perception pipeline by incorporating broader environmental and robot state information to enable richer and more appropriate responses. Finally, the identified differences between human-operated and AI-driven guidance should be leveraged to systematically improve and evaluate AI-driven guidance approaches.

7 Limitations

This study has several limitations that should be acknowledged. First, due to the AI-driven guidance's requirement for a quieter environment, noise levels differed between the human-operated and AI-driven conditions, potentially introducing confounding variables that could affect user perceptions and evaluations. Second, the collected data exhibited imbalance, with fewer AI-driven guidance sessions compared to human-operated sessions. This imbalance might have impacted the robustness and reliability of the comparative analyses. Third, the robot's navigation capability was restricted to predefined waypoints, limiting its ability to dynamically adapt to visitor behavior and potentially influencing the naturalness and interactivity of the provided guidance. Finally, the utterances analyzed in this study were transcribed using OpenAI Whisper-large-v3 instead of human annotators, which may have introduced transcription errors or noise.

8 Ethical Considerations

This study was approved by our institute's ethics committee and was conducted following relevant ethical guidelines. Informed consent was explicitly obtained from all participants prior to their involvement in the experiment. The participants in this study were visitors voluntarily attending the Nifrel facility, located in Osaka, Japan. All participants were native Japanese speakers and did not receive any monetary or material compensation for their participation. To ensure privacy and ethical standards, all collected data were anonymized, securely stored, and analyzed only in aggregate form, minimizing any potential risk to the participants. We used GPT-4o and GPT-4o-mini via OpenAI's API for dialogue generation and image recognition, respectively, and Google's Speech-to-Text API for speech recognition. All these artifacts were utilized in accordance with their respective terms of use and licenses.

9 Acknowledgments

This work was supported by the JST Moonshot R&D Grant number JPMJMS2011. We express our sincere gratitude to Professor Ogawa and his team at Nagoya University, Japan; to Professor Yoshikawa, Professor Iwasaki, Professor Meneses, and their colleagues at Osaka University, Japan; and to the staff at Nifrel for their invaluable cooperation in conducting the field experiment.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and Shyamal Anadkat. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Safa Jameel Al-Kamil and Róbert Szabolcsi. 2024. Optimizing path planning in mobile robot systems using motion capture technology. *Results in Engineering*, 22:102043–102052.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, and 1 others. 2022. RT-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*.

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, and 1 others. 2023. Do as I can, not as I say: Grounding language in robotic affordances. In *Proceedings of the 2023 Conference on Robot Learning*, pages 287–318.

Jian Chen, Chengshuai Wu, Guoqing Yu, Deepak Narang, and Yuexuan Wang. 2021. Path following of wheeled mobile robots using online-optimization-based guidance vector field. *IEEE/ASME Transactions on Mechatronics*, 26(4):1737–1744.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, and 3 others. 2023. PaLM-E: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Yi Fan, Jiquan Chen, Gabriela Shirkey, Ranjeet John, Susie R Wu, Hogeun Park, and Changliang Shao. 2016. Applications of structural equation modeling (sem) in ecological studies: an updated review. *Ecological Processes*, 5(1):1–12.

Ao Guo, Ryu Hirai, Atsumoto Ohashi, Yuya Chiba, Yuiko Tsunomori, and Ryuichiro Higashinaka. 2024. Personality prediction from task-oriented and open-domain human–machine dialogues. *Scientific Reports*, 14(1):3868–3879.

Ao Guo, Atsumoto Ohashi, Ryu Hirai, Yuya Chiba, Yuiko Tsunomori, and Ryuichiro Higashinaka. 2025. User personality and its influence on the performance of pipeline and end-to-end task-oriented dialogue systems. *Scientific Reports*, 15(1):20745–20763.

Chengyang Han and Baoying Li. 2023. Mobile robot path planning based on improved A* algorithm. In *Proceedings of the 11th IEEE Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, pages 672–676.

Hiroshi Ishiguro, Tetsuo Ono, Michita Imai, Takeshi Maeda, Takayuki Kanda, and Ryohei Nakatsu. 2001. Robovie: An interactive humanoid robot. *Industrial Robot: An International Journal*, 28(6):498–504.

Ross Jacobucci, Kevin J Grimm, and John J McArdle. 2016. Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 23(4):555–566.

Krzysztof Krejtz, Andrew Duchowski, Tomasz Szmidt, Izabela Krejtz, Fernando González Perilli, Ana Pires, Anna Vilaro, and Natalia Villalobos. 2015. Gaze transition entropy. *ACM Transactions on Applied Perception (TAP)*, 13(1):1–20.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Latural Language Processing*, pages 230–237.

Haokun Liu, Yaonan Zhu, Kenji Kato, Atsushi Tsukahara, Izumi Kondo, Tadayoshi Aoyama, and Yasuhisa Hasegawa. 2024. Enhancing the LLM-based robot manipulation through human-robot collaboration. *IEEE Robotics and Automation Letters*, pages 6904–6911.

Shota Mochizuki, Sanae Yamashita, Kazuyoshi Kawasaki, Reiko Yuasa, Tomonori Kubota, Kohei Ogawa, Jun Baba, and Ryuichiro Higashinaka. 2023. Investigating the intervention in parallel conversations. In *Proceedings of the 11th International Conference on Human-Agent Interaction*, pages 30–38.

Mikio Nakano, Yuji Hasegawa, Kazuhiro Nakadai, Takahiro Nakamura, Johane Takeuchi, Toyotaka Torii, Hiroshi Tsujino, Naoyuki Kanda, and Hiroshi G Okuno. 2005. A two-layer model for behavior and dialogue planning in conversational service robots. In *Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3329–3335.

Amit Kumar Pandey, Rodolphe Gelin, and AMPSH Robot. 2018. Pepper: The first machine of its kind. *IEEE Robotics & Automation Magazine*, 25(3):40–48.

Konstantinos I Roumeliotis and Nikolaos D Tselikas. 2023. ChatGPT and Open-AI models: A preliminary review. *Future Internet*, 15(6):192–216.

Kerstin Ruhland, Christopher E Peters, Sean Andrist, Jeremy B Badler, Norman I Badler, Michael Gleicher, Bilge Mutlu, and Rachel McDonnell. 2015. A review of eye gaze in virtual agents, social robotics and HCI: Behaviour generation, user interaction and perception. In *Proceedings of the 2015 Computer Graphics Forum*, pages 299–326.

Tim Schreiter, Lucas Morillo-Mendez, Ravi T Chadalavada, Andrey Rudenko, Erik Billing, Martin Magnusson, Kai O Arras, and Achim J Lilienthal. 2023. Advantages of multimodal versus verbal-only robot-to-human communication with an anthropomorphic robotic mock driver. In *Proceedings of the 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 293–300.

Philip Sedgwick. 2014. Spearman's rank correlation coefficient. *BMJ*, 349:1–4.

Dhruv Shah, Błażej Osiński, and Sergey Levine. 2023. LM-Nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Proceedings of the 2023 Conference on Robot Learning*, pages 492–504.

Ketan Rajshekhar Shahapure and Charles Nicholas. 2020. Cluster quality analysis using silhouette score. In *Proceedings of the 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 747–748.

Sebastian Thrun, Michael Beetz, Maren Bennewitz, Wolfram Burgard, Armin B Cremers, Frank Dellaert, Dieter Fox, Dirk Haehnel, Chuck Rosenberg, and Nicholas Roy. 2000. Probabilistic algorithms and the interactive museum tour-guide robot minerva. *The International Journal of Robotics Research*, 19(11):972–999.

Rudolph Triebel, Kai Arras, Rachid Alami, Lucas Beyer, Stefan Breuers, Raja Chatila, Mohamed Chetouani, Daniel Cremers, Vanessa Evers, and Michelangelo Fiore. 2016. Spencer: A socially aware service robot for passenger guidance and help in busy airports. *Field and Service Robotics*, pages 607–622.

Biel Piero E Alvarado Vásquez and Fernando Matía. 2020. A tour-guide robot: Moving towards interaction with humans. *Engineering Applications of Artificial Intelligence*, 88:103356–103373.

Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2023a. Realpersonachat: A realistic persona chat corpus with interlocutors'own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 852–861.

Sanae Yamashita, Shota Mochizuki, Kazuyoshi Kawasaki, Tomonori Kubota, Kohei Ogawa, Jun Baba, and Ryuichiro Higashinaka. 2023b. Investigating the effects of dialogue summarization on intervention in human-system collaborative dialogue. In *Proceedings of the 11th International Conference on Human-Agent Interaction*, pages 316–324.

Akishige Yuguchi, Seiya Kawano, Koichiro Yoshino, Carlos Toshinori Ishi, Yasutomo Kawanishi, Yutaka Nakamura, Takashi Minato, Yasuki Saito, and Michihiko Minoh. 2022. Butsukusa: A conversational mobile robot describing its own observations and internal states. In *Proceedings of the 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1114–1118.

Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S Yu. 2023. Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226*.

Kelly H Zou, Julia R Fielding, Stuart G Silverman, and Clare MC Tempany. 2003. Hypothesis testing I: Proportions. *Radiology*, 226(3):609–613.

A Prompt Structure for AI-driven Guidance

Section	Content
Instruction	You are "Teleco", a mobile guide robot in Nifrel, introducing pufferfish to visitors. Based on the provided information, generate utterances that continue the dialogue history naturally.
Guidance Outline	- Move sequentially between small and large tanks. - Explain Okinawa Pufferfish at the small tank (approx. 5 turns). - Introduce Mouse and Makeup Pufferfish at the large tank (approx. 10 turns). - Return to the starting position.
Rules	- Maintain a friendly and engaging tone. - Responses should be concise (max 50 characters). - Ensure smooth conversation flow without abrupt transitions. (omitted)
Animal Information	- **Okinawa Pufferfish**: Tropical species, first successfully bred in Nifrel, identified by two black bands and white spots. - **Mouse Pufferfish**: Largest species in the tank, expands body with spines when threatened. - **Makeup Pufferfish**: Named for its eye patterns resembling makeup, known as "Map Puffer" in English. (omitted)
Gaze Information	- Gaze Objects: Large Tank, Fish, Pufferfish - Gaze Description: There is a large tank with several fish swimming around.

Table 6: Prompt structure for the dialogue system (English translation of the original Japanese prompt).

B Operator's Post-dialogue Questionnaire

ID	Operator Questionnaire Items
Q1	I was able to remotely control the robot to effectively engage in conversation.
Q2	The visitor was satisfied with the interaction.
Q3	I felt that my communication was appropriate.
Q4	Through our conversation, I successfully deepened my interest in the visitor's exhibit.
Q5	The robot maintained an appropriate distance.
Q6	The robot was looking at the same exhibit as the visitor.
Q7	My speech matched the robot's actions.
Q8	I was able to operate the robot as intended using the controller.
Q9	The camera feed provided sufficient information for controlling the robot.
Q10	I effectively used commands to control the robot's movements and facial expressions.

Table 7: Operator's post-dialogue questionnaire (translated from Japanese).

C Distinctive 4-grams between Human-operated and AI-driven Guidance

Human-operated (JP)	Human-operated (EN)	AI-driven (JP)	AI-driven (EN)
ついてきて	Follow me	はあるかな	Is there any?
ハリセンボンの仲間な	Porcupinefish's companion	質問はあるか	Do you have any questions?
についてき	About this	でも聞いてね	Feel free to ask
実はハリセンボンの仲間	Actually, Porcupinefish's companion	なんでも聞いて	Ask me anything
けど実はハリセンボンの	But actually, Porcupinefish's companion	ななんでも聞い	Ask me anything
僕について	Follow me	あるかななん	Is there anything?
サンゴ礁に生息する	Inhabits coral reefs	かななんでも	Maybe anything
フグだよおち	It's a living pufferfish	か質問はある	Is there a question?
よおちょぼぐちと目	Pursed mouth and eyes	何か質問は	Any questions?
生息するフグだ	A pufferfish that inhabits	大型の水槽に	In a large aquarium

Table 8: Top 10 distinctive 4-grams between Human-operated and AI-driven guidance.

D Examples of Human-operated and AI-driven Dialogue

Human-operated	
Speaker	Utterance
Teleco	We’re moving, so follow me!
Teleco	This is an Okinawa pufferfish.
Teleco	It was successfully bred in Japan. Amazing, right?
Visitor	That’s amazing.
Teleco	Right? These were born in August 2022.
Teleco	When they grow up, they can reach about 30 cm in size.
AI-driven	
Speaker	Utterance
Teleco	In this tank, we have the Okinawa pufferfish! It’s a type of puffer that lives in tropical waters!
Teleco	Do you have any questions?
Visitor	Is it poisonous?
Teleco	There are a few reports of Okinawa pufferfish being poisonous, so it’s not suitable for eating. Do you have any other questions?
Teleco	Let’s move to the next location!
Teleco	In this large tank, we have two types of pufferfish: the Spot-fin porcupinefish and the Masked pufferfish! Can you tell the difference?

Table 9: Example of dialogue for human-operated and AI-driven guidance (translated from Japanese).