The Gap in the Strategy of Recovering Task Failure between GPT-4V and Humans in a Visual Dialogue

Ryosuke Oshima[1] **Seitaro Shinagawa**[2] **Shigeo Morishima**[3]
[1]Waseda University [2]Nara Institute of Science and Technology
[3]Waseda Research Institute for Science and Engineering
ryosukeoshima@fuji.waseda.jp, sei.shinagawa@is.naist.jp, shigeo@waseda.jp

Abstract

Goal-oriented dialogue systems interact with humans to accomplish specific tasks. However, sometimes these systems fail to establish a common ground with users, leading to task failures. In such cases, it is crucial not to just end with failure but to correct and recover the dialogue to turn it into a success for building a robust goal-oriented dialogue system. Effective recovery from task failures in a goal-oriented dialogue involves not only successful recovery but also accurately understanding the situation of the failed task to minimize unnecessary interactions and avoid frustrating the user. In this study, we analyze the capabilities of GPT-4V in recovering failure tasks by comparing its performance with that of humans using Guess What?! Game. The results show that GPT-4V employs less efficient recovery strategies, such as asking additional unnecessary questions, than humans. We also found that while humans can occasionally ask questions that doubt the accuracy of the interlocutor's answer during task recovery, GPT-4V lacks this capability.

1 Introduction

Goal-oriented dialogue systems work with humans on tasks to achieve a goal (de Vries et al., 2017; Kottur et al., 2021; Ma et al., 2022). They do not always succeed in their tasks in one shot due to the failure to establish a common ground of dialogue (Clark, 1996) with their human interlocutors. The task failure occurs due to various factors, including human error (Oshima et al., 2023), system error (Hudeček and Dusek, 2023; Mazuecos et al., 2021), and misunderstandings between the two (Paek and Horvitz, 2000).

In human-to-system dialogue, it is important for humans to finally achieve a successful goal regardless of the factor of task failure that occurs along the way. In this case, the system needs the capability to continue the failure dialogue and cooperatively recover from the task failure rather than terminating the dialogue (Benotti and Blackburn, 2021a). For example, suppose a task where an interactive autonomous driving system and a user tackle the task of going to an interior shop. The task may fail due to unexpected events, such as the destination being closed for construction or the user miscommunicating the desired location (Ma et al., 2022). In these cases, the system must offer alternatives or confirm the user's statements to recover from the task failure.

Of course, successful recovery from failure is not the only requirement for this dialogue task. As a goal-oriented dialogue, the system also demands minimizing the number of interactions to avoid frustrating the user. The system should have a dialogue strategy that makes good use of the information in the failure dialogue history and efficiently recovers the task.

While the recent Vision-Language Models (VLMs) integrated with Large Language Models (LLMs) have garnered attention for their ability to solve tasks at a high level through dialogue (OpenAI, 2024; Liu et al., 2023b), the performance of these VLMs in "failure task recovery" remains unclear. Investigating and analyzing these models' failure task recovery capabilities can lead to the development of robust dialogue systems for real-world applications. For example, instead of relying solely on VLM for the entire task recovery process, we can enhance the system's overall performance by implementing rule-based modules and preprocessing VLM inputs to compensate for VLM's weaknesses.

In this paper, we analyze the VLM's ability to recover the course of the dialogue as a first step toward the goal of building a system that can efficiently return to success after a task failure. We consider a problem setting in which the system performs a recovery action after a goal-oriented visual dialogue with a human interlocu-

Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 728–745
September 18–20, 2024. ©2024 Association for Computational Linguistics

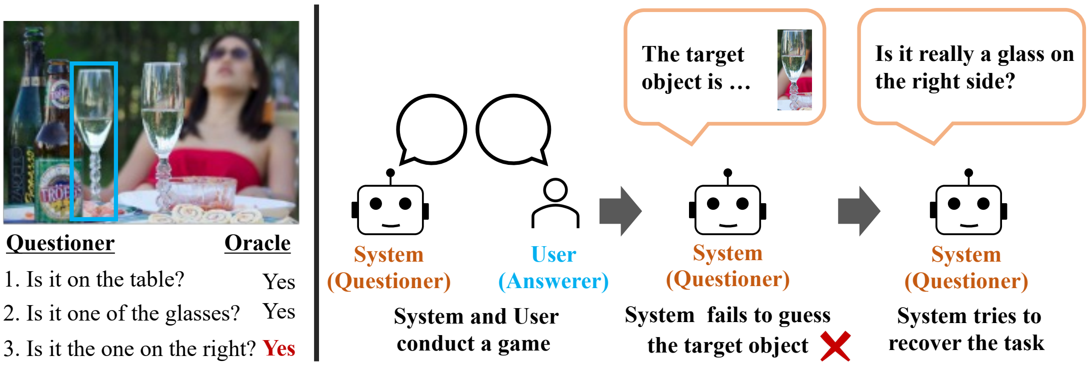


Figure 1: The *left* figure shows a failed game at Guess What?! Game by human annotators, which is included in Guess What?! Dataset (de Vries et al., 2017). The target object is outlined in blue. In this example, the questioner failed to correctly guess the target object due to the third wrong answer by the answerer. The *right* figure overviews the failure recovery task. A system and a user conducted the game, but the system guessed a different object. In this example, the system attempts to recover from a task failure to success by asking an additional question.

tor fails. Specifically, we leverage Guess What?! Game (de Vries et al., 2017), which is a widely used game of goal-oriented visual dialogue. Guess What?! Game (de Vries et al., 2017) is appropriate for a study of failure task recovery in dialogue because the goal of this game is building common ground directly.

In experiments, we prepare 100 task failure samples in Guess What?! Game (de Vries et al., 2017) and analyze the recovery capabilities of VLM by comparing them in human-to-system dialogue. Our experiments reveal that VLM struggles with recovering from task failure in a goal-oriented dialogue. It frequently performs unnecessary recoveries and uses ineffective repair utterances. Furthermore, humans tend to check the reliability of previous answers when errors are suspected. On the other hand, GPT-4V does not question prior utterances and fails to generate repair utterances that express doubt.

2 Related Work

2.1 Repair in Dialogue

Repair is one of the key interactional mechanisms to achieve shared understanding and coordination when miscommunication has occured (Schegloff et al., 1977; Schegloff, 1992; Purver et al., 2018). The construction of a robust dialogue system that can recognize and use repair has been discussed because this miscommunication also occurs in human-machine conversation (Purver et al., 2018; Balaraman et al., 2023; Shaikh et al., 2024). In visual goal-oriented dialogue research field, clarification requests has been mainly discussed as a repair utterance (Benotti and Blackburn, 2021b; Shi et al., 2022; Deng et al., 2023; Madureira and Schlangen, 2023; Chiyah-Garcia et al., 2023). Some research (Shi et al., 2022; Deng et al., 2023) deal with clarification questions for disambiguation in Minecraft game, where the system interacts with the user in the task of moving and building blocks according to the user's instructions. Chiyah-Garcia et al. (2023) used SIMMC dataset to analyze what information is important for a shopping assistant in a virtual shop to interpret a user's clarification requests. These studies focus on building a system that can perform or understand recovery "during" the dialogue to complete the task successfully the first time.

Although making clarification requests is a well-known dialogue strategy to avoid miscommunication (San-Segundo et al., 2001; Benotti and Blackburn, 2021b), it is hard for a system and humans to achieve a successful goal without failures all the time. In this paper, we consider the problem setting where once the dialogue is over and the task has ended in failure, how to turn it into a success. It is noted that the commonly used concept of "repair" or "repair utterance" in Conversational Analysis forms part of the recovery task and corresponds to Step 2 in the recovery flow introduced later (§3.2).

2.2 Recover in Tasks Other than Dialogue

Here, we describe previous works on addressing task failures and converting them into successes in non-dialogue tasks. Huang et al. (2022); Wang et al. (2023) worked on a task where a robot follows human instructions. When the robot failed to execute the instructions, Huang et al. (2022); Wang et al. (2023) utilized the LLM's strong reasoning abilities to correct the failures and achieve success. Huang et al. (2023); Fan et al. (2023); Zhang et al. (2023) focused on Automated Program recovery (APR), which aims to automatically fix software bugs and errors in programming.

These works focus on recovery that occurs solely within the systems. In contrast, dialogue recovery requires the systems to cooperatively interact with humans, presenting two main challenges. First, a system needs to minimize interactions to avoid frustrating the user. Second, the system must understand that human response errors cause task failure and not place too much trust in past dialogues (Oshima et al., 2023). Given the complexity of recovery tasks after a task failure, it is important to conduct a detailed analysis of VLMs' capabilities.

3 Guess What?! Game and Failure Recovery Task

3.1 Guess What?! Game

Guess What?! Game (de Vries et al., 2017) is a two-player game in which a questioner asks yes or no questions to identify a target object, and an answerer [1] answers those questions. We don't use other visual goal-oriented dialogue tasks (Kottur et al., 2021; Haber et al., 2019). This is because these dialogue scenarios involve too detailed object positioning within images (*e.g.*, "Do you like the second sweater from the right in the bottom row?") or require recognition of multiple images simultaneously, which VLMs generally perform poorly (Wu and Xie, 2023; Yang et al., 2023).

3.2 Failure Recovery Task Definiton

In this study, we consider the situation where the questioner and the answerer played a game, but the questioner failed to predict the target object (task failure). In such a case, the questioner should follow up with the user to ensure they can recover the task successfully from failure. We focus on this recovery process after task failure once this game is over. Figure 1 shows provides an overview of the recovery task. This recovery task requires a high success rate and emphasizes more efficient recovery. It is desirable to achieve recovery with as few additional questions as possible and, if feasible, without any additional questions.

[1] de Vries et al. (2017) calls an answerer "Oracle". Instead, we use "answerer" in this paper to avoid misunderstanding because an interlocutor can make mistakes and not always give a perfect answer.

Figure 2 shows a detailed flow of the questioner's (system side) recovery. The questioner can only choose one from two actions: 1) asking a question or 2) guessing a target object. The questioner performs the recovery in four steps.

Step 1: The questioner determines additional questions or re-prediction of objects based on information from the failure game. This step corresponds to the dialogue act classification module in goal-oriented dialogue. Given an image I, a dialogue history $H = (\underbrace{(Q_1, A_1)}_{H_1}, \ldots, \underbrace{(Q_{t-1}, A_{t-1})}_{H_{t-1}})$, and an object of failed prediction O_f, the questioner determines action A (a_1: asking an additional question or a_2: re-guessing the target object).

Step 2: If the questioner determines to ask an additional question (a_2), it asks the question Q_t according to the image I, the dialogue history H, and the object of the failed prediction O_f.

Step 3: The questioner judges if the target object could be uniquely determined by an additional question and the answerer's answer A_t. If the questioner judged unique, it proceeds to Step 4; if not, returns to Step 2. The dialogue history now contains additional questions and answers, $H \rightarrow H' = (\underbrace{(Q_1, A_1)}_{H_1}, \ldots, \underbrace{(Q_t, A_t)}_{H_t})$.

Step 4: If the questioner didn't ask any additional questions, it guesses the target object using the dialogue history H. If the questioner asks an additional question, it guesses the target object using the updated dialogue history H'.

These four steps are divided into two main parts: the decision to make a repair utterance and its actual execution (Steps 1 and 2) and the ability to correctly understand and process the repair utterance (Steps 3 and 4). In this study, we analyze the outcomes of recovery in Guess What?! Game (de Vries et al., 2017), where all four reasoning abilities are challenged at once in Section 5.1, and then we focus our analysis on the first two steps in Section 5.2.

4 Experiments

In this study, we analyze the success rate and features of the failure recovery task for humans and GPT-4V[2] (OpenAI, 2024). We first collect the failure game of Guess What?! (de Vries et al., 2017) to investigate failure recovery capability. Then, we

[2] We used the GPT-4 Turbo API through all the experiments in this study.

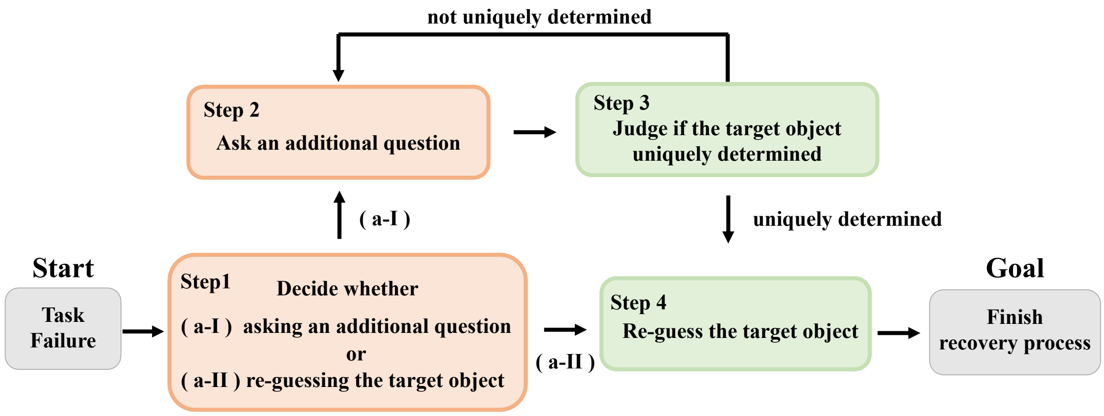


Figure 2: Questioner's recovery flow. The recovery process is divided into two main parts: Step1, 2 and Step3, 4.

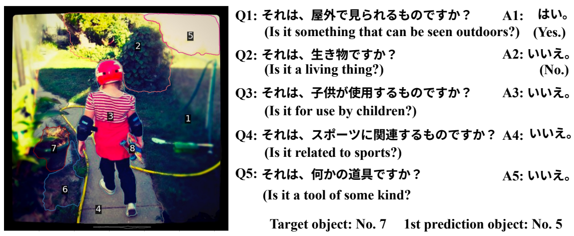


Figure 3: Failure game example

collect human and GPT-4V (OpenAI, 2024) failure task recoveries for the collected task failure game.

4.1 Failure Game Collection

We simulated human-to-system dialogues in Guess What?! Game using GPT-4V to collect data. Specifically, we set up GPT-4V as a questioner and answerer and collected failed games by having them play a game. While the ideal scenario would involve a human as the answerer, our preliminary experiments demonstrated that GPT-4V is sufficiently capable of playing Guess What?! Game. This led us to adopt the method of using GPT-4V in both roles for this study.

Figure 3 shows an example of a generated failure game. The questioner failed to narrow down the target object with five questions and could not guess it accurately. The following describes the details of the GPT-4V roles for the questioner and answerer (See Appendix A.1 for overview diagrams of these models).

Questioner's Role The questioner's role is divided into two parts: a model that makes questions (called a questioner model) and a model that guesses the target objects (called a guesser model), as de Vries et al. (2017)'s proposed baseline model. The questioner model inputs a game image and dialogue history and outputs a question. The guesser model takes an image with numbers assigned to objects by SoM (Yang et al., 2023) (called SoM-image) and dialogue history as inputs and outputs the number of the target object. We applied SoM to the input images of the guesser model because GPT-4V has better inference ability with number assignment images than with understanding Visual Prompt (Yang et al., 2023), and it is impossible to output a target object's bounding box[3].

Answerer's Role The answerer model takes SoM-image, a dialogue history, and the number of the correct object as inputs and outputs a yes/no answer.

Game Collection Details and Results According to (de Vries et al., 2017), the guesser model tried guessing the target object after the questioner and answerer models exchanged questions and answers five times. We sampled 815 pairs of images and target objects from the Guess What?! dataset's test data. Then, we excluded any target objects that were too small or positioned at the edges of the images, as recognizing these objects demands high image recognition capabilities beyond the scope of our study. As a result, we collected 100 failed games. The collected games include samples where GPT-4V, acting as the answerer, made errors, resulting in failed games. We also adopted these samples as examples that simulate actual human-to-system dialogues because humans can also make mistakes in their answers due to misinterpretations or unintentional mistakes (Oshima et al., 2023).

[3]When GPT4-V takes a prompt "Output the human bounding box." and an image as inputs, it returns an unreliable bounding box or says "I'm unable to directly output bounding boxes or any form of visual annotations".

4.2 Failure Task Recovery Collection

We collected samples on the recovery tasks performed by humans and GPT-4V (OpenAI, 2024), using the collected task-failed games in Section 4.1. Specifically, we conducted three experiments; *GPT-4V-all*, *GPT-4V-Q* , and *Human-all* experiment. Table 1 presents the relationships among these experiments. These experiments vary depending on who is responsible for each step of the questioner's recovery flow, which is introduced in Section 3.2. We describe the details of the three experiments below.

4.2.1 Human-all Experiment

We conducted an experiment to collect human recovery actions in failed games (called Human-all experiment). In collecting human recovery actions, two annotators each assumed the roles of questioner and answerer. The annotator in the questioner role worked with the answerer to address and recover from game failures, using the details of the failed task (game image, first predicted objects, and dialogue history from the failed game). The annotator in the answerer role received information about the details of the failed task and the correct target object. We created a demo application to collect humans' recovery actions. Humans were monitored to ensure they were not cheating and diligently working on tasks. There are other ways to collect data through crowdsourcing, but we did not employ them in this case because they are fraught with problems, such as using Large Language Models (Veselovsky et al., 2023).

The data collection had 12 native Japanese speaker participants, each performing recovery actions for 25 games. We assigned 25 game recovery tasks to each annotator using a collection of 100 failed games in Section 4.1. This means that three annotators worked on the recovery task for each game, resulting in 300 recovery samples collected in total.

4.2.2 GPT-4V-all Experiment

We also collected samples on recovery actions by GPT-4V. In this experiment, GPT-4V is responsible for all four steps. We prepared four GPT-4V models that perform each of the four steps of the recovery flow described in Section 3.2. We provided all GPT-4V models with the SoM-image, the number of the object predicted in the failed game, and the dialogue history as inputs (See Appendix A.2 for details of these models.). By comparing Human-all experiment to GPT-4V-all experiment, we can validate GPT-4V's ability to recover after task failure relative to human recovery ability.

	Step1	Step2	Step3	Step4
GPT-4V-all	GPT4	GPT4	GPT4	GPT4
GPT-4V-Q	human	GPT4	human	human
Human-all	human	human	human	human

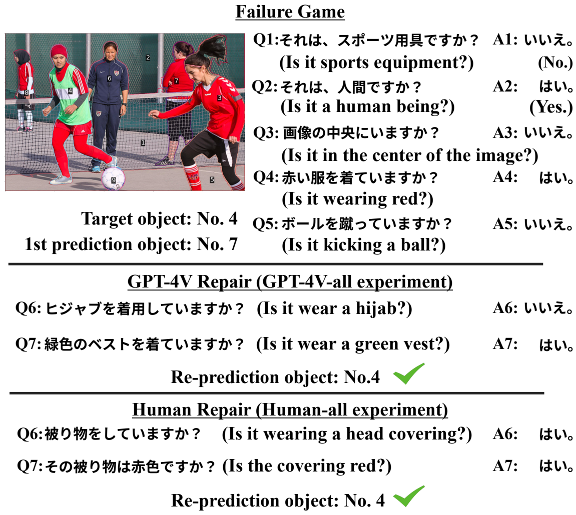
Table 1: Questioner's roles of humans and GPT-4V in each step in each experiment. Answer's roles were performed by humans at all experiments

4.2.3 GPT-4V-Q Experiment

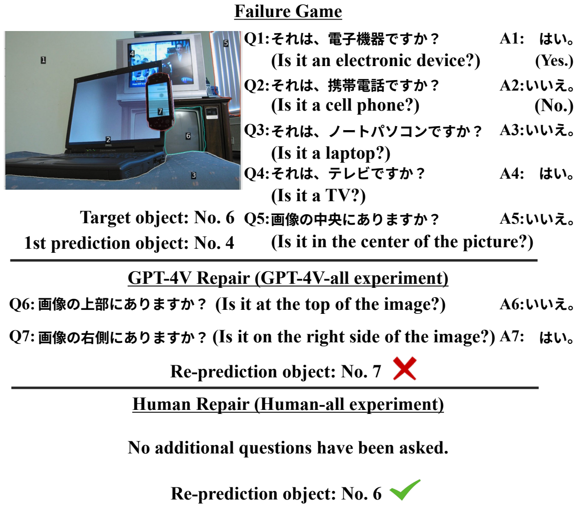
Then we also collected samples on recovery actions by GPT-4V and humans. In this experiment, GPT-4V is in charge of only step 2 among the four steps, and humans are in charge of the other steps. By comparing Human-all experiment to GPT-4V-Q experiment, it is possible to more directly assess the GPT-4V's ability to repair utterance (corresponding to step 2). In both GPT-4V-all experiment and GPT-4V-Q experiment, we collected three recovery samples per failed game, resulting in a total of 300 recovery samples, which is the same number as Human-all experiment. In all three experiments, if a human or GPT-4V repeated additional questions (Step 2) more than 11 times, we counted the attempt as a failure.

4.3 Evaluation

High recovery success rates and efficiency are essential for the failure recovery task, as mentioned in Section 3.2. Accordingly, we evaluate and analyze from the following two perspectives: (1) the success rate of recovery and (2) the average number of repair utterances required for a successful recovery. The success rate of recovery is defined as the number of recovered games over the total number of failed games. We define a successful recovery action as a successful re-prediction of the target object. We calculate the average number of repair utterances required for successful correction by averaging the additional questions made during the second step of the recovery process. This average includes instances where no repair utterances were necessary, counted as zero. A lower average indicates a more efficient recovery process, preferable for optimal outcomes.



(a) Example of the successful recovery of both the human and GPT-4V



(b) Example of the successful recovery of the human but the failure of GPT-4V

Figure 4: The recovery results for the same failed game. The image is labeled with numbers by SoM (Yang et al., 2023)). The upper proper dialogue represents the dialogue history during the task failure.

5 Results

5.1 Human vs GPT-4V for Recovery Action

We compared GPT-4V-all experiment and Human-all experiment to evaluate the abilities of humans and GPT-4V in implementing recovery actions after failing Guess What?! Game.

The success rate of recovery actions by GPT-4V is significantly lower than in humans (about 36.7% lower), which means that the failure task recovery in Guess What? Game (de Vries et al., 2017) is even difficult for GPT-4V. Figure 4a shows an example where both GPT-4V and the human was successful. GPT-4V successfully re-guessed the target object by asking two additional questions (Q6 and Q7), much like the human did, although using a different method of questioning. Figure 4b presents a case where GPT-4V failed, but the human succeeded. The human identified the target object without asking additional questions, whereas GPT-4V asked two questions (Q6 and Q7) and still failed to predict correctly. Despite confirming that the object was not a mobile phone in Q2, it incorrectly guessed the target object as No.7. This example shows failures in Step 1, 3 modules, which are responsible for deciding whether to ask an additional question, and in the Step 4 module, which is responsible for predicting the final object.

Next, we compared the efficiency of failure recovery tasks between humans and GPT-4V by analyzing the average number of repair utterances. As noted in Section 4.3, the recovery process must be efficient in human interactions. We calculated the average number of repair utterances only for successful cases because efficient recoveries are only relevant when the recovery task is successful (attempting a quick fix is pointless if it fails). Table 2 shows the average turn of repair utterances. The GPT-4V-all experiment is more than twice utterances as many as the Human-all experiment, indicating a less efficient recovery strategy in GPT4-V compared to humans.

5.2 First Half Recovery Steps Analysis

In this section, we focus on GPT-4V's ability to decide and successfully execute repair utterances (steps 1 and 2) rather than just understanding and using them (steps 3 and 4).

5.2.1 Step1: Deciding Recovery Action

We compared the actions chosen by GPT-4V with those selected by humans in step 1. Specifically, for each game, we tallied and compared the number of times actions (a-I) asking an additional question and (a-II) re-guessing the target object were chosen. We selected actions chosen at least twice by the human and GPT-4V across three recovery tasks for the same failure sample as the actions by the recovery executor.

Table 3 compares the actions selected by humans (Human-all experiment) and GPT-4V (GPT-4V-all experiment). As Section 5.1 indicates, human re-

	Success rate	Average turn
GPT-4V-all	50.0%	2.43
GPT-4V-Q	74.7%	2.00
Human-all	86.7%	1.13

Table 2: Success rate of recovery actions and the number of repair utterances (step 2) in each experiment.

	GPT4-V (a-I)	GPT4-V (a-Ⅱ)
Human (a-I)	65	5
Human (a-Ⅱ)	28	2

Table 3: The number of actions selected by Humans and GPT-4V (a-I or a-Ⅱ). Diagonal elements show the number of times Humans and GPT-4V made the same selections. Note that these counts are from Step 1 of GPT-4V-all experiment and Human-all experiment.

covery actions have a high success rate and are a strong baseline. Thus, GPT-4V should choose actions that are similar to those chosen by humans in most cases. Table 3 shows that GPT-4V selects about 67% of the same actions as humans, and GPT-4V often opts to select (a-I) action even in cases where humans choose (a-II) action. This result demonstrates that GPT4-V fails to provide efficient questions and choose speedy recovery actions. This behavior is undesirable because making efficient failure task recovery is crucial in goal-oriented dialogues with humans.

5.2.2 Step2: Asking an Additional Question (Repair Utterance)

Next, we analyzed the repair utterances from Step 2. Specifically, we compared GPT-4V-Q experiment, in which humans handled all steps except Step 2, with Human-all experiment, in which humans were responsible for all steps.

Table 2 shows the results for task success rates and the number of repair utterances. When comparing GPT-4V-Q experiment to Human-all experiment, we observe that only replacing Step 2 with GPT-4V results in a 12% decrease in success rate and an increase of 0.87 in average turns. Furthermore, in GPT-4V-all experiment where GPT-4V handles all steps, the success rate drops by an additional 24.7%, and the average turns increase by 0.43. This suggests that GPT-4V's impact in Step 2 contributes more to the increase in the number of turns than modules of other steps, which means that GPT-4V's repair utterances tend to include unnecessary questions.

Next, we analyzed the intents behind the utterances to compare the nature of repair utterances made by humans and GPT-4V. We asked humans

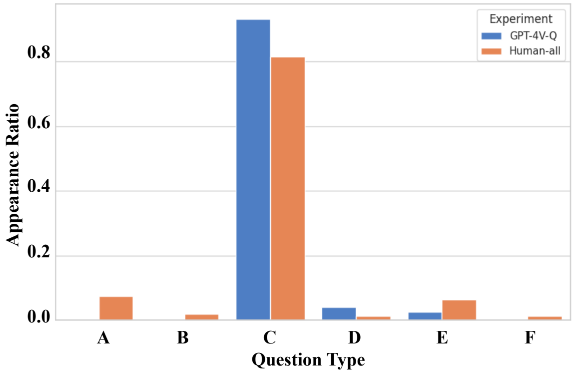


Figure 5: Question type distribution. We normalize by dividing the number of each question type by the total number of questions in each experiment because the number of questions differs between the two experiments.

and GPT-4V to select the intents behind their questions from six options and compared these selections. We assumed that GPT-4V could understand the intentions behind its questions, so we had GPT-4V select the intent of the questions. We also asked for explanations behind the selected options, and the lead author checked to see if the reason was plausible because GPT-4V does not always produce accurate outputs. We conducted preliminary experiments and prepared the following six types of questions (A)-(F) (see Appendix B for detailed explanations of question types.):

(A) The question that addresses the same object with different expressions: This questioning style is employed when there is a suspicion of inconsistencies or errors in the user's answers.

(B) The question with more or less the same meaning as the question during dialogue: This questioning style is used when there are suspected inconsistencies or errors in the user's answers.

(C) The question that proposes a hypothesis to narrow down the object in question: This type of question is used when there are no apparent errors or contradictions in the user's answers.

(D) The question that clarifies ambiguities in a previous question: It clarifies the context or perspective of the previous question.

(E) The question for confirmation, in case the object has already been narrowed down.

(F) Others. (In this case, we ask the annotator and GPT-4V to describe the question's intent in text form.)

Figure 5 shows the distribution of the intentions behind the questions asked by humans and GPT-4V. First, more type (C) questions exist in both Human-

Failure Game

Q1: Is it sporting goods? No.
Q2: Is it on snow? No.
Q3: Is it a vehicle? No.
Q4: Related to winter sports? Yes.
Q5: Is it a ski lift? No.
Target :No. 4 , 1st prediction: No. 7

GPT-4V-Q experiment

Q6(C): Is that a ski board? No.	Q11(C): Is it a goggle? No.
Q7(C): Is it a snowboard? No.	Q12(C): Is it a ski pole? No.
Q8(C): Is it ice skates? No.	Q13(C): Are those ski boots? No.
Q9(C): Is it gloves? No.	Q14(C): Is it a ski lift seat? No.
Q10(C): Is it a hat? No.	Q15(C): Is it a ski lift prop? No.

Human-all experiment

Q6(C): Is that something you put on your feet? No.
Q7(C): Is it something you carry on your back? No.
Q8(A): Is it a living thing? Yes.
Q9(C): Is it a person on the right? Yes.

Figure 6: Example of a human asking a type (A) question, whereas GPT-4V does not. Q6 (C) indicates that the first additional question is the intent of type (C). The dialogue conducted in Japanese is translated into English.

all and GPT-4V-all experiments. This is because many samples in the failed game set, such as the example in Figure 4, require additional questions to narrow down the objects. A significant difference between the human-only experiment and the GPT-4V with human experiment is that the humans can ask many non-type (C) questions. Specifically, humans ask type (A) or (B) questions about 9.3% of the time, whereas GPT-4V rarely asks these types of questions. This indicates that GPT-4V cannot recognize or doubt mistakes and inconsistencies based on the user's input, leading it to focus predominantly on questions that narrow down objects.

Figure 6 shows an example where a human asks a type (A) question while GPT-4V does not. In this example, the human suspects an error in the answer and attempts to correct the course of the dialogue by asking, "Is it a living thing?" (Question type A). In contrast, GPT-4V likely overtrusts the response, "Related to winter sports? Yes," and continues to ask questions focused on objects related to winter sports (Question type C). As a result, GPT-4V fails to correct the course of the dialogue and cannot identify the target object within ten questions, leading to an unsuccessful recovery task. These results indicate that GPT-4V fails to recognize or question erroneous responses and tends to blindly trust the user's input.

6 Discussion

We observed that GPT-4V is significantly poor at the failure recovery task (§5.1), and GPT-4V's approach differs from the strong baseline of human behavior in both Step 1 and Step 2 (§5.2). The significant difference in failure task recovery capabilities between humans and GPT-4V can be attributed to the models' difficulty with logical reasoning (Creswell et al., 2023; Pan et al., 2023; You et al., 2023). GPT-4V may struggle to integrate three pieces of information from failed games (game image, dialogue text, and first prediction object) to identify potential target objects. Unlike typical goal-oriented dialogues, conducting failure task recovery requires understanding complex dialogue and game situations. Therefore, the preprocessing step that explicitly organizes the context of the failed game and dialogue rather than executing direct recovery actions may be practical. In Guess What?! Game, output which objects remain as potential targets is an example of this strategy.

We also found that GPT-4V tends to refrain from questioning the interlocutor's answer (§5.2.2). This feature is undesirable for the failure task recovery in goal-oriented dialogues, where the user's answers might contain errors (Oshima et al., 2023). This issue is not crucial during initial task attempts because humans do not frequently make response errors. However, when the task fails, the possibility of user answer errors increases. Therefore, considering the possibility of user errors is a key factor when developing recovery strategies. If a system cannot doubt the user's answers, it may fail to correct errors or waste time, as shown in Figure 6. To address this issue, instructing the LLM first to evaluate the correctness of the user's answers and then use this evaluation to guide the recovery action may be effective.

One future direction to use VLMs (Liu et al., 2023b,a) as a recovery model is rethinking model training. For example, creating synthetic datasets that include incorrect utterances and using them for instruction-tuning data. This approach allows VLMs to explicitly learn from erroneous scenarios, potentially enhancing their abilities to recover task failure accurately.

7 Limitations

In this study, we examined the failure recovery task in Guess What?! Game, where one speaker only responds with "Yes." or "No." However, this research does not address the recovery capabilities of GPT-4V and humans in more complex goal-orientated dialogues like autonomous driving dialogue systems. Our results may differ for languages other than Japanese, so it is essential to analyze GPT-4V's recovery performance in English, its most proficient language. This study focuses on GPT-4V, raising concerns about the generalizability of our findings to other vision-language models. We tested the failure recovery task with LLaVA-1.5 (Liu et al., 2023a), but it did not recover the tasks adequately, which suggests that the recovery task would require capabilities comparable to GPT-4V.

We are concerned about the method of collecting intents by directly asking humans or using GPT-4V. This method assumes a causal relationship between subjective reasoning and actual behavior. As Ayaß (2015) recommended, it is preferable to analyze speech intentions based on objective actual behavior rather than subjective reasoning.

8 Conclusion

We tackled the failure recovery task in Guess What?! Game and analyzed GPT-4V's capabilities. The results showed that GPT-4V demonstrated a significantly lower ability to correct task failures than humans. Furthermore, GPT-4V tended to perform unnecessary repair utterances, ask inefficient questions, and fail to doubt users' answers. In future work, we aim to investigate the generalizability of our findings to real-world goal-oriented dialogues.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP21H05054, JP21K17806, JP24H00742, and JP24H00742.

References

Ruth Ayaß. 2015. Doing data: The status of transcripts in conversation analysis. *Discourse Studies*, 17(5):505–528.

Vevake Balaraman, Arash Eshghi, Ioannis Konstas, and Ioannis Papaioannou. 2023. No that's not what I meant: Handling third position repair in conversational question answering. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 562–571, Prague, Czechia. Association for Computational Linguistics.

Luciana Benotti and Patrick Blackburn. 2021a. Grounding as a collaborative process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 515–531, Online. Association for Computational Linguistics.

Luciana Benotti and Patrick Blackburn. 2021b. A recipe for annotating grounded clarifications. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4065–4077, Online. Association for Computational Linguistics.

Javier Chiyah-Garcia, Alessandro Suglia, Arash Eshghi, and Helen Hastie. 2023. 'what are you referring to?' evaluating the ability of multi-modal dialogue models to process clarificational exchanges. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 175–182, Prague, Czechia. Association for Computational Linguistics.

Herbert Clark. 1996. *Using Language*. Cambridge University Press.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2023. Selection-inference: Exploiting large language models for interpretable logical reasoning. In *The Eleventh International Conference on Learning Representations*.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1080–1089. Institute of Electrical and Electronics Engineers.

Yang Deng, Shuaiyi Li, and Wai Lam. 2023. Learning to ask clarification questions with spatial reasoning. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2113–2117, New York, NY, USA. Association for Computing Machinery.

Zhiyu Fan, Xiang Gao, Martin Mirchev, Abhik Roychoudhury, and Shin Hwei Tan. 2023. Automated repair of programs from large language models. In *Proceedings of the 45th International Conference on Software Engineering*, ICSE '23, page 1469–1481. IEEE Press.

Janosch Haber, Tim Baumgärtner, Ece Takmaz, Lieke Gelderloos, Elia Bruni, and Raquel Fernández. 2019. The PhotoBook dataset: Building common ground through visually-grounded dialogue. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1895–1910, Florence, Italy. Association for Computational Linguistics.

Kai Huang, Xiangxin Meng, Jian Zhang, Yang Liu, Wenjie Wang, Shuhao Li, and Yuqing Zhang. 2023. An empirical study on fine-tuning large language models of code for automated program repair. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 1162–1174.

Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. 2022. Inner monologue: Embodied reasoning through planning with language models. In *arXiv preprint arXiv:2207.05608*.

Vojtěch Hudeček and Ondrej Dusek. 2023. Are large language models all you need for task-oriented dialogue? In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–228, Prague, Czechia. Association for Computational Linguistics.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ziqiao Ma, Benjamin VanDerPloeg, Cristian-Paul Bara, Yidong Huang, Eui-In Kim, Felix Gervits, Matthew Marge, and Joyce Chai. 2022. DOROTHIE: Spoken dialogue for handling unexpected situations in interactive autonomous driving agents. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4800–4822, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Brielen Madureira and David Schlangen. 2023. Instruction clarification requests in multimodal collaborative dialogue games: Tasks, and an analysis of the CoDraw dataset. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2303–2319, Dubrovnik, Croatia. Association for Computational Linguistics.

Mauricio Mazuecos, Patrick Blackburn, and Luciana Benotti. 2021. The impact of answers in referential visual dialog. In *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pages 8–13, Gothenburg, Sweden. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Ryosuke Oshima, Seitaro Shinagawa, Hideki Tsunashima, Qi Feng, and Shigeo Morishima. 2023. Pointing out human answer mistakes in a goal-oriented visual dialogue. In *VISION-AND-LANGUAGE ALGORITHMIC REASONING Workshop in International Conference on Computer Vision 2023*.

Tim Paek and Eric Horvitz. 2000. Conversation as action under uncertainty. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, UAI'00, page 455–464, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Liangming Pan, Alon Albalak, Xinyi Wang, and William Wang. 2023. Logic-LM: Empowering large language models with symbolic solvers for faithful logical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3806–3824, Singapore. Association for Computational Linguistics.

Matthew Purver, Julian Hough, and Christine Howes. 2018. Computational models of miscommunication phenomena. *Topics in Cognitive Science*, 10(2):425–451.

Ruben San-Segundo, Juan Manuel Montero, and Jose Manuel Pardo. 2001. Designing confirmation mechanisms and error recover techniques in a railway information system for Spanish. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

E.A. Schegloff. 1992. Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American Journal of Sociology*, 97(5):1295–1345.

E.A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

Omar Shaikh, Kristina Gligorić, Ashna Khetan, Matthias Gerstgrasser, Diyi Yang, and Dan Jurafsky. 2024. Grounding gaps in language model generations. *Preprint*, arXiv:2311.09144.

Zhengxiang Shi, Yue Feng, and Aldo Lipani. 2022. Learning to execute actions or ask clarification questions. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2060–2070, Seattle, United States. Association for Computational Linguistics.

Veniamin Veselovsky, Manoel Horta Ribeiro, and Robert West. 2023. Artificial artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. *Preprint*, arXiv:2306.07899.

Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. 2023. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *arXiv preprint arXiv:2302.01560*.

Penghao Wu and Saining Xie. 2023. V*: Guided visual search as a core mechanism in multimodal llms. *arXiv preprint arXiv:2312.14135*.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *Preprint*, arXiv:2310.11441.

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. 2023. IdealGPT: Iteratively decomposing vision and language reasoning via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11289–11303, Singapore. Association for Computational Linguistics.

Quanjun Zhang, Chunrong Fang, Yuxiang Ma, Weisong Sun, and Zhenyu Chen. 2023. A survey of learning-based automated program repair. *ACM Trans. Softw. Eng. Methodol.*, 33(2).

A GPT-4V Model Details

A.1 Models for Failure Game Collection

Figure 7 shows an overview of the model used for failure game collection in Section 4.1. We provide the original SoM-image to the questioner and guesser models, while the answerer model receives an SoM-image with the target object highlighted in a yellow frame. Table 4 shows the text prompts provided to the questioner and guesser models. Table 5 indicates the prompts given to the answerer model.

A.2 Models for Failure Task Recovery Collection

The basic framework of the models is the same as the GPT-4V model prepared in failure game collection. Each model uses the SoM-image but with different text prompts. The text prompts consist of two main parts: the system prompt, which is the rule of Guess What?! Game, and the user prompt, which is the specific instructions performed by GPT-4V. While the system prompt is consistent across all models, the user prompt varies by step. Table 6 illustrates the system prompt, and Table 7 and 8 provide examples of the user prompt.

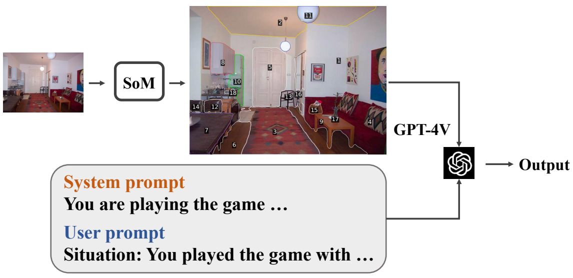


Figure 7: Overview of the model used for the failure game collection (§4.1) and failure task recovery collection (§4.2).

B Details of Question's intentions

Table 9 shows a detailed description of each question type. We gave the same content written in Table 9 to GPT-4V while we showed the users a Japanese translation of the content described in the table. Table 10 presents cases where humans selected type (F) for the question intent, along with descriptions of their intent.

Prompt type	Example
Questioner prompt	Instruction: You are now playing the following game with a user. Rule of the game: You are trying to guess the target object in the image by asking questions. User will answer yes or no to the question about the user's target object. Note that target objects can also include living creatures such as humans and animals. Generate a question "in Japanese" to guess the target object. 例：「それは、人間ですか？」、「それは、画像の左側にありますか？」、「それは、野菜ですか？」など、なんでも構いません。 (Example: You can ask anything, such as “Is that a human?” or “Is that a vegetable?” or “Is it on the left side of the picture?”) Dialogue history: 質問:「それは、テニスラケットですか？」 回答: はい。 では、質問を生成してください。 (Now, please make a question.) 質問： (Question:)
Guesser prompt	Instruction: You are now playing the following game with a user. Rule of the game: You are trying to guess the target object in the image by asking you questions. User will answer yes or no to the question about the user's target object. You are trying to guess the target object through dialogue. Dialogue history: 質問: それは、テニスラケットですか？ 回答: はい。 質問: それは、人間ですか？ 回答: いいえ。 質問: それは、画像の右側にありますか？ 回答: はい。 質問: それは、男性ですか？ 回答: いいえ。 質問: それは、テニスコートですか？ 回答: いいえ。 Please predict the target object number. Be sure to focus your prediction on one number! Number:

Table 4: Questioner’s role prompt examples for failure game collection (§4.1).

Prompt type	Example
Answerer prompt	Instruction: You are now playing the following game with a user. Rule of the game: The user is trying to guess the target object in the image by asking you questions. Answer yes or no to the question about the user's target object. The target object is labeled with number 8, surrounded by yellow box and its category is TENNIS RACKET. 質問には、「はい。」または「いいえ。」で答えてください。 (Please answer with “Yes” or “No”.) 質問：それは、人間ですか？ (Question: Is it a human?) 回答： (Answerer:)

Table 5: Answerer’s role prompt examples for failure game collection (§4.1).

Prompt type	Example
System prompt	You are now playing the following game with a user. You are a professional in this game. Rule of the game: You are trying to guess the target object in the image by asking questions. User will answer yes or no to the question about the user's target object. Note that numbered objects are candidates for target objects and target objects can also include living creatures such as humans and animals.

Table 6: System prompt for the failure task recovery collection (§4.2)

Prompt type	Example
User prompt of step1	Situation: You played the game with the user and also predicted the object. However, the object you predicted (number 3) was not the right target object, and the task failed. Therefore, you need to take a repair action to turn this game into a success (guessing the correct target object) instead of a failure. This repair action can be either asking additional questions or re-predicting the object. Which is better? Dialogue when you fail to predict the target object (Dialogue before the first prediction): 質問：それはスポーツ用品ですか？ 回答：いいえ。 質問：それは生き物ですか？ 回答：はい。 質問：それは人間ですか？ 回答：はい。 質問：それは、画像の中央にいますか？ 回答：いいえ。 質問：それは、画像の右側にいますか？ 回答：いいえ。 Instruction: Please answer the number of the action you take. Note that target objects must be numbered. (1). You do not ask additional questions and re-guess the target object (2). You ask additional questions in order to re-guess the target object
User prompt of step2	Situation: You played the game with the user and also predicted the object. However, the object you predicted (number 3) was not the right target object, and the task failed. Dialogue when you fail to predict the target object (Dialogue before the first prediction): 質問：それはスポーツ用品ですか？ 回答：いいえ。 質問：それは生き物ですか？ 回答：はい。 質問：それは人間ですか？ 回答：はい。 質問：それは、画像の中央にいますか？ 回答：いいえ。 質問：それは、画像の右側にいますか？ 回答：いいえ。 Instruction: You have determined that you need to ask additional questions. Please make one question to re guess the target object based on failed dialogue. Notes: 1. Numbered objects are candidates for a target object. 2. Please do not ask additional questions using the number assigned to the object or ask questions that mention that number or letter! では、質問を生成してください。 (Now, please make a question.) 質問： (Question:)

Table 7: Prompts of Steps 1 and 2 for the failure task recovery collection (§4.2)

Prompt type	Example
User prompt of step3	Situation: You played the game with the user and also predicted the object. However, the object you predicted (number 1) was not the right target object, and the task failed. Then, you asked additional questions to the user in order to re-guess the target object. Dialogue when you fail to predict the target object (Dialogue before the first prediction): 質問：それは、鏡の中に映っていますか？ 回答：いいえ。 質問：それは、洗面台ですか？ 回答：はい。 質問：それは、人間ですか？ 回答：いいえ。 質問：それは、タイルでできていますか？ 回答：いいえ。 質問：それは、水道の蛇口ですか？ 回答：いいえ。 Additional questions and answers to re-predict the correct target object (Dialogue after the first prediction): 質問：それは、壁に取り付けられていますか？ 回答：いいえ。 Instruction: Please review the previous conversation and decide if the target object in this game has been clearly identified. Note that the object should be identifiable by a number. Respond with: (1) The object has been clearly identified. (2) The object has not been identified, and further questions are necessary.
User prompt of step4	Situation: You played the game with the user and also predicted the object. However, the object you predicted (number 3) was not the right target object, and the task failed. Dialogue when you fail to predict the target object (Dialogue before the first prediction): 質問：それはスポーツ用品ですか？ 回答：いいえ。 質問：それは生き物ですか？ 回答：はい。 質問：それは人間ですか？ 回答：はい。 質問：それは、画像の中央にいますか？ 回答：いいえ。 質問：それは、画像の右側にいますか？ 回答：いいえ。 Additional questions and answers to re-predict the correct target object (Dialogue after the first prediction): 質問：それは画像の左側にいますか？ 回答：はい。 Instruction: Please read the dialogue history above and re-predict the target object number. Be sure to focus your prediction on one number! The target object: Number

Table 8: Prompts of Steps 3 and 4 for the failure task recovery collection (§4.2)

Question intention	**Description**
(A) The question that addresses the same object with different expressions.	This questioning style is employed when there is a suspicion of inconsistencies or errors in the user's answers. It involves exploring the same object in an image through various expressions. This method helps identify any inconsistencies or errors in the user's answers by exploring different aspects of the same object and examining the object from multiple angles.
(B) The question with more or less the same meaning as the question during dialogue.	This questioning style is used when there are suspected inconsistencies or errors in the user's answers. It involves slight rephrasing of previous questions using similar terms to clarify and rectify any misunderstandings. Example: Rephrase the question "Are people using it?" in the dialogue history as "Is it something that people are holding?" or "Is it a human figure?" as "Is it really a human figure?"
(C) The question that proposes a hypothesis to narrow down the object in question.	This type of question is used when there are no apparent errors or contradictions in the user's answers. It introduces a hypothesis to further refine and specify the inquiry, aiming to deepen the exploration of the object in question. This approach helps gather more precise information about the object being discussed.
(D) The question that clarifies ambiguities in a previous question	Employed in cases where the meaning of the previous question has caused ambiguity, this questioning style seeks to align the understanding between the questioner and the respondent. It clarifies the context or perspective of the previous question. It clarifies the context or perspective of the previous question, as in the example: "Is it on the left?" might be followed by a clarifying question, "Is it on the left side of the image?" to specify the context.
(E) The question for confirmation, in case the object has already been narrowed down	This questioning style is utilized when the answers so far have no detected errors or contradictions. It aims to reaffirm the accuracy and certainty of the information provided by the respondent, ensuring a solid and shared understanding of the discussed object or situation.
(F) Others.	

Table 9: Question intentions and detail description.

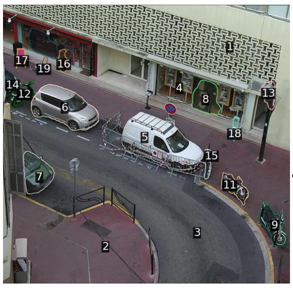
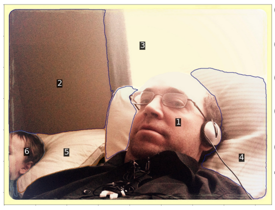
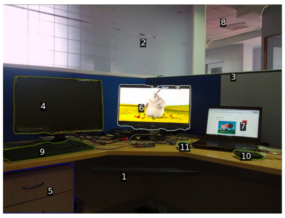
Recovery cases	Description of question intentions
 **Q1: Is it a vehicle? A1: Yes.** **Q2: Is it a four-wheeled vehicle? A2: Yes.** **Q3: Is it white? A3: Yes.** **Q4: Does it have a load on the roof? A4: No.** **Q5: Is it in the center of the image? A5: No.** ----------------------------------- **Q6: Is that a motorcycle? A6: No.** **Q7: Is it green? A7: No.** **Q8: Is it fenced? A8: Yes.** **Target object: No. 5, 1st prediction object: No. 6, Repair prediction object: No. 5** ✓	Q6: (F) In the dialogue, I noticed errors or contradictions in the answers and, while trusting the question "Is it a vehicle?" with the response "Yes," restarted the discussion about other topics to explore the possibility of other objects. Q7: (C) Q8: (A)
 **Q1: Is it someone wearing headphones? A1: No.** **Q2: Is it a person wearing glasses? A2: Yes.** **Q3: Is the person lying down? A3: Yes.** **Q4: Is the person on the bed? A4: Yes.** **Q5: Is the person wearing headphones? A5: No.** ----------------------------------- **Q6: Is it on the left side of the screen? A6: Yes.** **Target object: No. 6, 1st prediction object: No. 5, Repair prediction object: No. 6** ✓	Q6: (F) Since it was uncertain whether the baby was wearing glasses, I asked additional questions to make sure.
 **Q1: Is that the character you see on the screen? A1: No.** **Q2: Is it an electronic device? A2: Yes.** **Q3: Is it on a desk? A3: Yes.** **Q4: Is it the monitor in the center of A4: Yes.** **Q5: Is it on the right side of the image? A5: No.** ----------------------------------- **Q6: Is it the monitor on the left side of the image? A6: Yes.** **Target object: No. 4, 1st prediction object: No. 6, Repair prediction object: No. 4** ✓	Q6: (F) Questions to notice mistakes or inconsistencies in the answers during the dialogue and to narrow the object based on them.
 **Q1: Is it in the kitchen? A1: Yes.** **Q2: Does it use water? A2: No.** **Q3: Is it used for cooking? A3: Yes.** **Q4: Does it use heat? A4: Yes.** **Q5: Is it an oven? A5: No.** ----------------------------------- **Q6: It on the right side of the screen? A6: Yes.** **Target object: No. 7, 1st prediction object: No. 6, Repair prediction object: No. 7** ✓	Q6: (F) Questions that narrow down objects based on noticing errors or contradictions in responses during a conversation.

Table 10: The samples and the description where a human selected the question type (F).