Pointing out Human Answer Mistakes in a Goal-Oriented Visual Dialogue

Ryosuke Oshima[1] Seitaro Shinagawa[2] Hideki Tsunashima[1] Qi Feng[1] Shigeo Morishima[3]
[1]Waseda University [2]Nara Institute of Science and Technology
[3]Waseda Research Institute for Science and Engineering
{ryosukeoshima@fuji, h.tsunashima@asagi, fengqi@ruri}.waseda.jp, sei.shinagawa@is.naist.jp, shigeo@waseda.jp

Abstract

Effective communication between humans and intelligent agents has promising applications for solving complex problems. One such approach is visual dialogue, which leverages multimodal context to assist humans. However, real-world scenarios occasionally involve human mistakes, which can cause intelligent agents to fail. While most prior research assumes perfect answers from human interlocutors, we focus on a setting where the agent points out unintentional mistakes for the interlocutor to review, better reflecting real-world situations. In this paper, we show that human answer mistakes depend on question type and QA turn in the visual dialogue by analyzing a previously unused data collection of human mistakes. We demonstrate the effectiveness of those factors for the model's accuracy in a pointing-human-mistake task through experiments using a simple MLP model and a Visual Language Model.

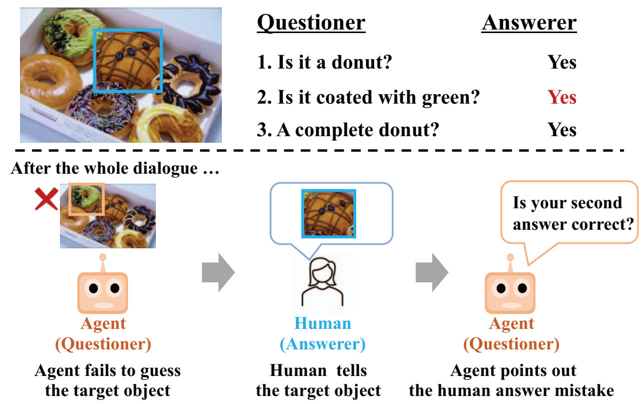

Figure 1. Overview of the pointing-human-mistake task. When the human interlocutor (Answerer) provides an incorrect answer (**red**) about the target object (**blue**) in a given image, the agent (Questioner) identifies and flags the mistake as a classification task. This process is illustrated in the bottom image.

1. Introduction

The communication between humans and intelligent agents has gained increasing attention due to its potential to solve various problems, such as making reservations and navigation. To further enhance this capability, visual dialogue, which utilizes multimodal context, has emerged as a promising approach to assist humans [18].

In real-world scenarios, human interlocutors may not always respond accurately to agents due to misinterpretations or unintentional mistakes. In fact, human-to-human visual dialogue data collection includes failed tasks with incorrect human answers. However, many current visual dialogue systems assume that the responses provided by the interlocutor are always correct [12, 17].

To address the potential impact of human mistakes on visual dialogue accuracy, we introduce a task where an agent identifies and points out mistakes made by a human interlocutor in response to the agent's questions. This task draws inspiration from Guess What?! [4], a well-established benchmark for evaluating the performance of agents in visual dialogue tasks. We extend the original task by adding steps for the agent to point out mistakes to the human, who can then acknowledge and review their errors.

Previous work [15] examines situations where the interlocutor intentionally provides incorrect answers, emphasizing uncooperative behavior. However, encountering uncooperative dialogue partners is rare in typical problem-solving scenarios. In our study, we initiate goal-oriented dialogues with the assumption that the interlocutor is cooperative, providing an opportunity to resolve the task. Unlike [15], our focus is on a scenario where the agent points out incorrect answers unintentionally provided by the interlocutor, which better reflects the real-world applications of visual dialogue.

The problem setting of pointing out mistakes shares similarities with the problem setting of clarification requests for uncertain answers during dialogue [2, 11]. This approach proves valuable as it allows an agent to ask questions about ambiguous human statements, thereby mitigating the risk of failure. However, it is important to note that even with this questioning strategy, complete mistake prevention is not guaranteed, as other factors may contribute to task failures. The problem setting of pointing out mistakes can be positioned as a problem of how to recover after a mistake has occurred and the task has failed, and this study aims to analyze the problem setting for this purpose. In this paper, we focus on a pointing-human-mistake task, distinct from traditional clarification requests, which allows us to explore the potential of agents in identifying and pointing out human answer mistakes. Leveraging the widely used Guess What?! Dataset [5], we construct Human Mistake Dataset to facilitate our investigation. Through a two-fold analysis, we establish clear correlations between the two key features and the occurrence of human mistakes: QA turns and Question types. Having identified the factors, we extend our study and implement them in a simple MLP model and a Visual Language Model (VLM). Through experiments, we demonstrate that incorporating these features leads to remarkable improvements in the agent's prediction performance, specifically in pointing out human mistakes during visual dialogues.

By providing insights into the importance of identifying human answer mistakes for enhancing multimodal communication in visual dialogue systems, our work offers valuable information for more effective and accurate visual dialogues in future applications.

2. Human Answer Mistake Analysis

2.1. The Pointing-human-mistakes Task

Guess What?! (Figure 1 top) is a two-player game in which a Questioner asks yes/no questions to identify a target object, and an Answerer[1] provides answers to those questions. In the pointing out human mistake task, an agent takes on the role of the Questioner, and a human interlocutor acts as the Answerer. The agent's goal is to point out mistakes made by the human interlocutor during the dialogue. Figure 1 bottom provides an overview of the pointing-human-mistake task. By using this task, we can investigate an agent's ability to recognize and address human mistakes in a multi-turn conversational setting accurately.

2.2. Human Mistake Dataset Construction

To analyze the features of human answer mistakes and facilitate learning-based methods, we constructed a dataset based on Guess What?! [4], which consists of human-to-human dialogues including three types of collections: *Success*, *Failure*, and *Incomplete*. We bring attention to the *Failure* collection that has never been used in ordinary visual dialogue settings. We constructed Human Mistake Dataset by randomly selecting 2,300 dialogues from the *Failure* collection, filtering out noisy samples with small target object sizes, following [5]. As a result, Human Mistake Dataset contains 365 dialogues, enabling us to analyze human answer mistakes in visual dialogue settings.

[1] [4] calls an interlocutor Oracle. Instead, we use Answerer in this paper because an interlocutor can make mistakes and not give a perfect answer.

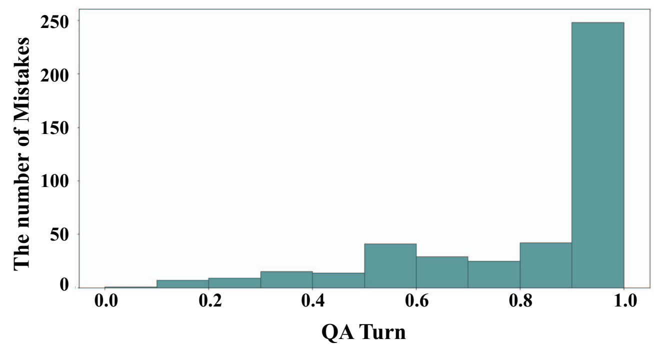

Figure 2. Histogram of relationship with QA turn and answer mistakes. The horizontal axis is normalized.

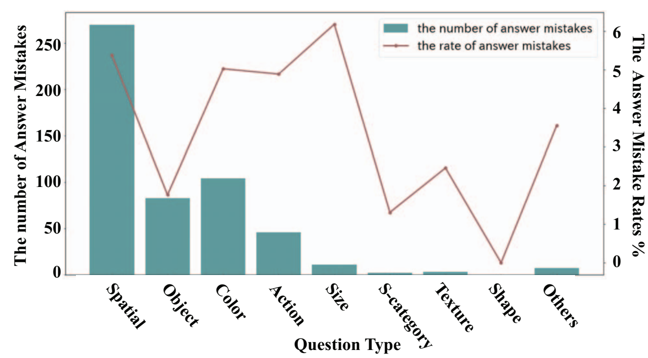

Figure 3. The mistake question rate. S-category means Super-category type. It is essential to note that the rates represent the mistake rates for each question type; therefore, the sum of all the rates does not equal 100%.

2.3. Analysis Methodology

Firstly, we examine the relationship between the QA turn and the occurrence of answer mistakes. We hypothesize that answer mistakes are more likely to occur as the dialogue progresses, since subsequent turns tend to involve more complex objects or relations to determine the target object. Secondly, we explore the relationship between the question type and the occurrence of answer mistakes. This aspect is particularly relevant given the findings of [16] and [14], who reported that the correct answer rate of Answerer models varies depending on the question type. By analyzing the relationship between question type and answer mistakes, we aim to gain insights into the nature of human answer mistakes in visual dialogue settings.

2.4. Analysis Results

QA Turn Figure 2 illustrates the relationship between the number of answer mistakes and the turn number in the dialogues. The QA turn (horizontal axis) is normalized as

$\frac{\text{current turn}}{\text{total turns}}$, which represents a relative turn in the dialogue because the number of turns varies in dialogues. As expected, the frequency of answer mistakes increases as the dialogue progresses, particularly in the latter half of the dialogue. Notably, 231 out of 431 samples of answer mistakes occurred in the last turn. (See Appendix C for details)

Question Type We investigated the impact of question types on human answer mistakes. We hypothesized that answer mistakes vary by the question types and conducted Fisher's exact test (refer to Appendix A and B for question types and details). Our findings showed that at least one question type has a significantly higher mistake rate than others, which supports our hypothesis.

Figure 3 illustrates the answer mistake rates for each question type. We calculated the mistake rates by dividing the number of incorrect answers by the total number of answers for each question type. The results indicate that humans are more likely to make mistakes when answering Spatial, Color, Action, and Size questions (See Appendix B for details.

3. Model

We employ two types of models and observe how effectively our findings in Section 2.4 improve the accuracy of the mistake-pointing out model.

3.1. MLP Model

Model Details We extend the Answerer model proposed in [4] by adding a classification head for detecting mistakes in human answers (Figure 4). We set a threshold of 0.5 for the model's output, allowing it to perform binary classification on whether the answer is correct or incorrect. We call this model the baseline model. While there are many possible methods to incorporate Question type and QA turn into the baseline model, we opt for a straightforward approach of introducing a new input to the baseline model (details provided in Appendix G). We call these models the Question type model and the QA turn model, respectively.

Pretraining Strategy Human mistakes in answers during dialogue are rare, making it costly to collect many human answer mistakes as it requires reviewing the entire dialogue and finding mistakes afterward. To overcome this challenge, we introduce pretraining with Synthetic Dataset using the *Success* collection of Guess What?! [4]. Specifically, we randomly flipped Yes/No in human answers and built a dataset of approximately 131k dialogues. Preliminary experiments showed that this pretraining strategy was effective (details are shown in Appendix D). To prevent the model from learning an unnatural dialogue flow in Synthetic Dataset (details are shown in Appendix E), we did not deliberately include the dialogue history as an input to the model in this study.

Training Details For pre-training, we use 70% of the Synthetic Dataset for training and 15% for validation, similar

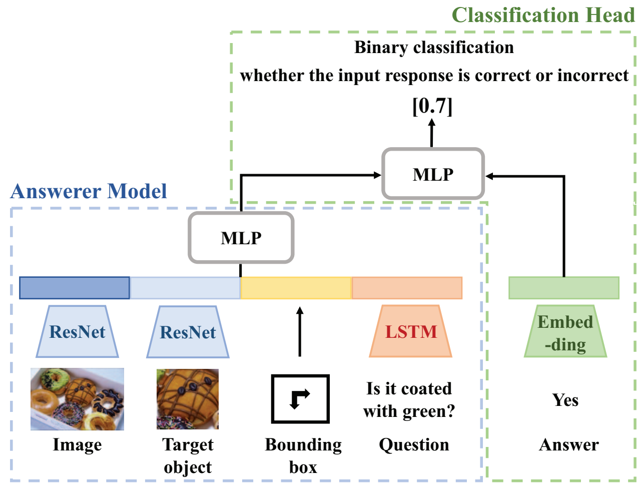

Figure 4. Baseline model overview diagram

to [4]. We omit the remaining 15% of the test dataset as the models are not evaluated during pre-training. We perform k-fold cross-validation with $k = 4$ while fine-tuning on Human Mistake Dataset. The results in Section 5.1 represent the average results for each experiment across the k cross-validation iterations. We set the learning rate to 10^{-4} for pre-training and 10^{-5} for fine-tuning with the Adam [9] optimizer.

3.2. Visual Language Model (VLM)

We further explore VLMs integrated with large language models, which have recently achieved high performance without task-specific fine-tuning [6]. We examine a representative VLM, OpenFlamingo[2] [1], in in-context learning setting to investigate VLMs' ability to point out human mistakes with prompts. Despite testing other VLM models (InstructBLIP [3] and BLIP2 [10]), OpenFlamingo outperformed them in our experiment (see Appendix H for details).

4. Experiment

In this study, we conduct an experiment to explore how incorporating the mistake tendency identified in Section 2.4 can improve the models' accuracy.

Dataset Our study uses two datasets: Synthetic Dataset for pre-training and Human Mistake Dataset for fine-tuning and evaluation. Human Mistake Dataset is divided into two parts: the same image dataset using the same images as the training data and the different image dataset with different images. It is noticed that both datasets are unknown to the dialogue text. We use both the same and different image datasets for evaluating MLP models and the different image dataset for evaluating a VLM (see Appendix I in detail).

[2]https://github.com/mlfoundations/open_flamingo

Evaluation To evaluate the performance of the model in pointing out answer mistakes, we use the F-score, Recall, and Precision which can evaluate imbalanced data (see Appendix F). Human Mistake Dataset used for evaluation is imbalanced; the number of positive cases (answer mistakes) is smaller than that of negative cases (correct answers). This is due to the rare occurrence of answer mistakes.

5. Results

5.1. MLP Models

We tested the baseline model, QA turn model, and Question type model to see if these additional features could improve the model's ability to point out mistakes in human answers based on the insights gained from our analysis in Section 2.4. We oversampled[3] samples in the fine-tuning stage to ensure an equal number of samples for each question type. Table 1 shows the results of MLP models' experiments.

The baseline model performed best in the same image dataset, whereas the QA turn model and Question type model outperformed it in the different image dataset. However, we were concerned that the accuracy of pointing out mistakes might decrease for responses other than the last one. Therefore, we evaluated the accuracy of the QA turn model both at the last time and at all other times. Table 2 shows that the QA turn model improved the F-score for the last response in the different image dataset without a significant decrease for other responses. The results indicate that in the *same* image dataset, including the additional input from the QA turn and Question type models led to unnecessary complexity, resulting in a lower F-score than the baseline model, which already captures visual information in known images. Conversely, in the *different* image dataset, the baseline model was less effective in capturing visual information, making the input from the QA turn and Question type models valuable. Furthermore, Table 2 demonstrates that the QA turn model may serve as a useful predictor instead of solely relying on learning the data distribution.

5.2. Visual Language Model

To assess the effectiveness of the human mistake features identified in Section 2.4, we also conducted tests using a VLM, which has recently been actively studied for its high accuracy. In particular, we used OpenFlamingo [1] and provided three types of prompts as input: (1) a normal prompt to point out human answer mistakes; (2) a question type hint prompt that gives information that humans are more likely to make mistakes in their responses to some question types; (3) a QA turn hint prompt to provide information that humans tend to make more mistakes as the dialogue progresses. We only conducted experiments with the different image dataset for OpenFlamingo. We also experimented with the case in which OpenFlamingo got a dialogue history. The reasons for these choices are provided in Appendix H. It is worth noting that in-context learning is known to be highly sensitive to few-shot prompting [13, 19], and we have not been able to exhaustively examine all possible prompts.

[3]A term used in the field of imbalanced data, and refers to the process of increasing the number of samples in categories with fewer instances.

Model	Same image	Different image
Baseline	**0.811**	0.482
QA turn	0.718	**0.514**
Question type	**0.743**	**0.527**

Table 1. The results of MLP models. From the top, the results are for the baseline, the QA turn model, and the Question type model. The score is F-score. The best score is in **black bold**, and the second-best score is in **blue**.

Model	Last time		Other than last time	
	Same	Different	Same	Different
Baseline	**0.875**	0.548	**0.714**	0.406
QA turn	0.789	**0.608**	0.609	0.406

Table 2. F-scores at the last time and at other times for the baseline model and QA turn model. Same and Different means the same image dataset and the different image dataset, respectively.

Prompt type	Without History	With history
Normal	0.313	0.325
QA turn hint	**0.374**	**0.377**
Question type hint	**0.366**	**0.372**

Table 3. The results of OpenFlamingo. The top three lines are the result of not including the dialogue history in the prompt, and the bottom three lines are the result of including the dialogue history in the prompt.

Table 3 shows the results with OpenFlamingo [1]. OpenFlamingo's F-score is lower than 38%, worse than the MLP models. It indicates that pointing out human answer mistakes seems difficult for OpenFlamingo without task-specific fine-tuning, unlike other typical V&L tasks such as VQA. We found that OpenFlamingo with Question type and QA Turn improved prediction accuracy. Human features are also effective for OpenFlamingo, whether or not dialogue history is considered.

6. Conclusion

In this paper, We considered a task in which agents point out human answer mistakes and analyzed what would be the key factors of human answer mistakes. We observed that mistakes were more common towards the end of the dialogue and varied based on question types. By incorporating these human mistake features, we enhanced the performance of both the MLP model and the VLM in our experiments on the actual pointing-human-mistake task. In future work, we aim to explore the generalizability of our findings to other goal-oriented visual dialogues.

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

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 23716–23736. Curran Associates, Inc., 2022.

[2] Luciana Benotti and Patrick Blackburn. 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, June 2021. Association for Computational Linguistics.

[3] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.

[4] Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 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, 2017.

[5] Tianai Dong, Alberto Testoni, Luciana Benotti, and Raffaella Bernardi. Visually grounded follow-up questions: a dataset of spatial questions which require dialogue history. In *Proceedings of Second International Combined Workshop on Spatial Language Understanding and Grounded Communication for Robotics*, pages 22–31, Online, Aug. 2021. Association for Computational Linguistics.

[6] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 2016 Conference on Computer Vision and Pattern Recognition*, pages 770–778. Institute of Electrical and Electronics Engineers, 2015.

[8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9:1735–1780, 1997.

[9] Diederik P Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *Proceedings of the 2015 International Conference on Learning Representations*, pages 1–15, 2015.

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 23–29 Jul 2023.

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

[12] Shoya Matsumori, Kosuke Shingyouchi, Yuki Abe, Yosuke Fukuchi, Komei Sugiura, and Michita Imai. Unified questioner transformer for descriptive question generation in goal-oriented visual dialogue. In *Proceedings of the 2021 International Conference on Computer Vision*, pages 1878–1887. Institute of Electrical and Electronics Engineers, 2021.

[13] Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 11054–11070. Curran Associates, Inc., 2021.

[14] Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. Beyond task success: A closer look at jointly learning to see, ask, and guesswhat. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, pages 2578–2587, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

[15] Anthony Sicilia, Tristan Maidment, Pat Healy, and Malihe Alikhani. Modeling non-cooperative dialogue: Theoretical and empirical insights. *Transactions of the Association for Computational Linguistics*, 10:1084–1102, 2022.

[16] Alberto Testoni, Claudio Greco, Tobias Bianchi, Mauricio Mazuecos, Agata Marcante, Luciana Benotti, and Raffaella Bernardi. They are not all alike: Answering different spatial questions requires different grounding strategies. In *Proceedings of the 2020 International Workshop on Spatial Language Understanding*, pages 29–38, Online, 2020. Association for Computational Linguistics.

[17] Tao Tu, Qing Ping, Govind Thattai, Gokhan Tur, and Prem Natarajan. Learning better visual dialog agents with pretrained visual-linguistic representation. In *Proceedings of the 2021 Conference on Computer Vision and Pattern Recognition*, pages 5618–5627. Institute of Electrical and Electronics Engineers, 2021.

[18] Hao Wang, Bin Guo, Yating Zeng, Yasan Ding, Chen Qiu, Ying Zhang, Lina Yao, and Zhiwen Yu. Enabling harmonious human-machine interaction with visual-context augmented dialogue system: A review. *arXiv preprint arXiv:2207.00782*, 2022.

[19] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR, 18–24 Jul 2021.

[20] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023.