# Looking for Confirmations:
# An Effective and Human-Like Visual Dialogue Strategy

**Alberto Testoni**
DISI - University of Trento
Trento - Italy
alberto.testoni@unitn.it

**Raffaella Bernardi**
CIMeC and DISI - University of Trento
Rovereto (TN) - Italy
raffaella.bernardi@unitn.it

## Abstract

Generating goal-oriented questions in Visual Dialogue tasks is a challenging and long-standing problem. State-Of-The-Art systems are shown to generate questions that, although grammatically correct, often lack an effective strategy and sound unnatural to humans. Inspired by the cognitive literature on information search and cross-situational word learning, we design *Confirm-it*, a model based on a beam search re-ranking algorithm that guides an effective goal-oriented strategy by asking questions that confirm the model's conjecture about the referent. We take the GuessWhat?! game as a case-study. We show that dialogues generated by *Confirm-it* are more natural and effective than beam search decoding without re-ranking.

## 1 Introduction

Quite important progress has been made on multimodal conversational systems thanks to the introduction of the Encoder-Decoder framework (Sutskever et al., 2014). The success of these systems can be measured by evaluating them on task-oriented referential games. Despite the high task-success achieved and the apparent linguistic well-formedness of the single questions, the quality of the generated dialogues, according to surface-level features, have been shown to be poor; this holds for systems based on both greedy and beam search (e.g. Shekhar et al. (2019); Zarrieß and Schlangen (2018); Murahari et al. (2019)). Testoni and Bernardi (2021a) found that when taking these surface-level features as a proxy of linguistic quality, the latter does not correlate with task success, and the authors point to the importance of studying deeper features of dialogue structures. We aim to develop a multimodal model able to generate dialogues that resemble human dialogue strategies.

Cognitive studies show that humans do not always act as "rational" agents. When referring to ob-
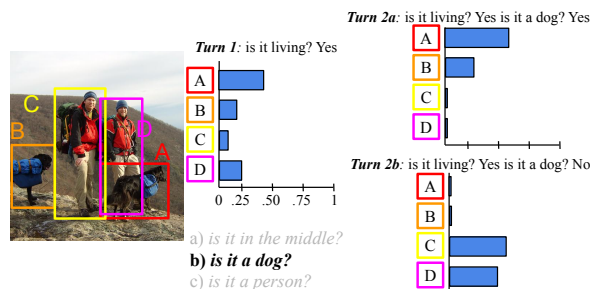


Figure 1: At turn 2, among the questions proposed by the beam search, *Confirm-it* chooses $b$ since it is the most suitable one to confirm the current conjecture.

jects, they tend to be overspecific and prefer properties irrespectively of their utility for identifying the referent (Gatt et al., 2013); when searching for information or when learning a language, they tend to follow confirmation-driven strategies. Modelling such behaviour in language learning, Medina et al. (2011) and Trueswell et al. (2013) propose a procedure in which a single hypothesized word-referent pair is maintained across learning instances, and it is abandoned only if the subsequent instance fails to confirm the pairing. Inspired by these theories, we propose a model, *Confirm-it*, which generates questions driven by the agent's confirmation bias.

Take the example of a referential guessing game in which an agent has to ask questions to guess an object in a given image. *Confirm-it* will ask questions that reinforce its beliefs about which is the target object, till proven otherwise. For instance, in Figure 1, after learning that the target is a living entity (turn 1), the agent conjectures the target is the dog on the right of the picture (though in principle, it could have been any of the candidates). Hence, the decoder generates the question that would let it confirm such belief, "is it a dog?". If its expectations are not met (viz., it receives a negative answer to such question - turn 2b), it moves its attention to another candidate object. We do not claim that our choice represents the optimal strategy to play the

9330

game, but we believe that it makes the generated dialogue more human-like.

To evaluate this strategy, we take as a test-bed GuessWhat?! (de Vries et al., 2017), a two-player game between a Questioner that has to guess the target, and an Oracle (called "external Oracle" in the following) who is aware of the target. The widely used architecture of the Questioner, GDSE, jointly trains a Question Generator (QGen) and a Guesser (Shekhar et al., 2019). We augment this architecture with a module that simulates an internal Oracle. Being an "internal" Oracle, at test time this agent does not know what the target object is: while at training time it learns to answer questions by receiving the gold standard datapoint (the question, the actual target, and the human answer), at test time it assumes the target is the candidate object to which the Guesser assigns the highest probability. Hence, the three modules of the Questioner straightforwardly cooperate one another. The internal Oracle guides the QGen to ask questions that reinforce the Guesser's beliefs. Concretely, at training time, through Supervised Learning (SL) the QGen learns to ask human-like questions turn-by-turn, the internal Oracle to answer them, and the Guesser to guess the target object once the dialogue ends. At test time, we implement a beam search re-ranking algorithm that simulates the single-conjecture learning strategy used by humans: among the questions the QGen generates via beam search, the algorithm promotes the questions whose answer (obtained via the internal Oracle that receives the candidate with the highest probability as the target) increases the most the model's confidence in its hypothesis about the target.

We run both quantitative and qualitative analyses, and evaluate the effectiveness of the dialogue strategy by asking human annotators to guess the target object given the dialogues generated by *Confirm-it*. We compare results giving the dialogue generated by *Confirm-it* when using the re-ranking algorithm and when generating the question proposed by the plain beam search. We show that the task accuracy of both the conversational agent and human subjects increases when receiving the dialogues generated by the *Confirm-it* re-ranking algorithm.

## 2 Related Work

For open-ended language generation, Holtzman et al. (2020) claim that decoding strategies that op-

timize for output with high probability (like beam search) lead to highly deteriorated texts, since the highest scores are often assigned to generic, incoherent, and repetitive sequences. Several works propose reranking strategies on the set of hypotheses produced by the beam search following different criteria (Dušek and Jurčíček, 2016; Blain et al., 2017; Agarwal et al., 2018; Borgeaud and Emerson, 2020; Hargreaves et al., 2021) to improve both the performance on a given task and the quality of the output. In this work, we present a cognitively-inspired reranking technique for a visual dialogue questioner agent.

In visual dialogue systems, the quality of the output has been improved mainly by aiming at reducing repetitions in the output. This goal has been achieved through Reinforcement Learning by adding auxiliary objective functions (Murahari et al., 2019), intermediate rewards (Zhang et al., 2018), regularized information gain techniques (Shukla et al., 2019), or intermediate probabilities with an attention mechanism (Pang and Wang, 2020). Different from these works, we do not use the Reinforcement Learning paradigm and, instead of focusing on improving surface-level features, we indirectly operate on the dialogue structure.

Ruggeri and Lombrozo (2015) studied the way children and young adults search for information while asking yes-no questions given a set of candidate hypotheses. The authors found that when prior knowledge favours some hypotheses over others, participants asked more hypothesis-scanning questions (i.e., questions that are tentative solutions, with a specific hypothesis that is directly tested). This is in line with the observation in Baron (2000) that humans phrase questions to receive an affirmative answer that supports their theory, and with the broader finding in Wason (1960) that they tend to select the information that is in accord with their prior beliefs. Inspired by these studies, we propose a new dialogue strategy for playing referential guessing games by exploiting the probabilities assigned by the Guesser module to different candidate objects.

## 3 Task and Dataset

GuessWhat?! (de Vries et al., 2017) is an asymmetric game involving two human participants who see a real-world image from MSCOCO (Lin et al., 2014). One of the participants (the Oracle) is secretly assigned a target object in the image, while

the other participant (the Questioner) has to guess it by asking binary (Yes/No) questions to the Oracle. The GuessWhat?! dataset consists of more than 150k human-human English dialogues containing on average 5.3 questions per dialogue.

## 4   Model and Re-ranking Strategy

Our model, *Confirm-it*, builds on GDSE (Shekhar et al., 2019). In the latter, the hidden state representation produced by a multimodal encoder is used to jointly train the question generator (QGen) and the Guesser module. The image is encoded with a ResNet-152 network (He et al., 2016) and the dialogue history is encoded via an LSTM network. QGen uses greedy search to generate questions. To this multi-tasking setting, *Confirm-it* adds an internal Oracle trained to answer the question at each turn. Moreover, it relies on beam search and, at inference time, it goes through a re-ranking phase which simulates the single-conjecture learning strategy. The model architecture is provided in the Supplementary Material (SM) and the algorithm is spelled out below.

---

**Algorithm 1** The *Confirm-it* algorithm

---
**Require:** History $H$, Beam size $B$, Max turns $T$,
     Image $I$, Distractors $D_{1:N-1}$, target $o_t$,
**Require:** Candidates $C_{1:N} \leftarrow D_{1:N-1} + o_t$
**Require:** Internal Oracle $IO$
**Require:** Target-aware external $Oracle$
 1: **for** $turn = 1 : T$ **do**
 2:    $\overset{\wedge}{p}(c_{k_{1:N}}) \leftarrow Guesser(H, I, C_{1:N})$
 3:    $c_h \leftarrow argmax(\overset{\wedge}{p}(c_{k_{1:N}}))$
 4:    $q_{1:B} \leftarrow QGen(H, I)$
 5:    $a^*_{1:B} \leftarrow IO(H + q_{1:B}, c_h)$
 6:    $H^*_{1:B} = H + (q_{1:B}, a^*_{1:B})$
 7:    $p^* \leftarrow Guesser(H^*_{1:B}, I, C_{1:N})$
 8:    $Q \leftarrow q_{1:B}[argmax(p^*(c_h))]$
 9:    $Oracle$ provides an answer $A$ to $Q$
 10:    $H \leftarrow H + Q + A$
 11: **end for**

---

Algorithm 1 describes the beam search re-ranking algorithm used by *Confirm-it* to promote the generation of an effective dialogue strategy. Given an image, a set of candidate objects, a target object $o_t$, and a beam size of $B$, at each dialogue turn the model predicts a probability distribution over the set of candidate objects given the current dialogue history. The candidate that receives the highest probability is considered the model's hy-

| Decoding Strategy | Accuracy 5Q |
|---|---|
| *Confirm-it* | 51.38 |
| Beam Search | 47.03 |
| Random Re-Rank | 46.29 |
| Greedy Search | 46.58 |

Table 1: Accuracy results of *Confirm-it* on the Guess-What?! test set when generating 5 questions per dialogue following different decoding strategies. *Confirm-it* refers to the accuracy achieved by our re-ranking algorithm compare to a random re-ranking of candidate questions, plain beam search, and greedy search.

pothesis $c_h$. The QGen outputs $B$ questions, ordered by their probability. Each of these questions is answered by the model's internal Oracle that receives $c_h$ as the target object. Among these $B$ questions, *Confirm-it* selects the question $Q$ that, paired with the answer provided by the internal Oracle, increases the most the model's confidence over $c_h$, measured as the probability assigned by the Guesser. The external Oracle (who is aware of the real target object $o_t$) answers $Q$, and this new question-answer pair is appended to the dialogue history. In SM we provide a step-by-step example of how *Confirm-it* works.

None of the features of our case-study are crucial for the method to be applied to other tasks, e.g. it does need the questions to be polar, it does not need the questions to be visually grounded, it does not need the dialogue to be asymmetrical.

**Implementation details**[1]   For the multi-task training, we adopt the *modulo-n* training proposed in Shekhar et al. (2019), i.e. we train the Oracle and guesser modules every $n$ (=7) epochs of the QGen. At inference time, we use a beam size of 3 and let the model generate dialogues of 5 turns.

## 5   Results

We study to what extent the re-ranking phase lets the model generate more effective and more natural dialogues. To this end, we evaluate the *Confirm-it* task-accuracy with and without the ranking phase[2] and report qualitative analyses of the dialogues.

**Task-accuracy**   Table 1 shows the task accuracy results of our model in the GuessWhat?! game.

---
[1] Our code is available at: https://github.com/albertotestoni/confirm_it

[2] Remember that *Confirm-it* is based on the GDSE architecture. For comparison, the accuracy reached by GDSE is 45.55% with greedy search and 46.40% using beam search decoding.

| | Human Accuracy ↑ | % games with repetitions ↓ | CHAIR-s ↓ | % Yes Last Turn↑ | % novel $(q_{t-1}, q_t)$ per dialogue ↑ |
|---|---|---|---|---|---|
| Beam Search | 70.8 | 38.50 | 31.07 | 71.87 | 36.75 |
| *Confirm-it* | 77.0 | 30.36 | 28.46 | 76.68 | 48.25 |
| Human Dialogues | 96.0 | 0.01 | 7.45 | 86.64 | - |

Table 2: *Human Accuracy* refers to the task accuracy achieved by human annotators when receiving dialogues generated by the plain Beam Search, *Confirm-it* re-ranking, or the original dialogues produced by human players from the GuessWhat?! test set. The other columns report relevant statistics of the dialogues: percentage of games with at least one repeated question verbatim, hallucination rate (CHAIR-s), percentage of positive answers in the final turn (% yes Last Turn), and percentage of consecutive questions not seen at training time (lexical overlap, % novel $q_{t-1}, q_t$ per dialogue).



**Human:**
Is it an utensil? <yes>
Is it the pizza cutter? <yes>

**Beam Search:**          **Confirm-it:**
is it food ? <no>          is it food ? <no>
is it a glass ? <no>       is it a utensil ? <yes>
is it a table ? <no>       is it on the left ? <yes>
is it a chair ? <no>       is it a fork ? <no>
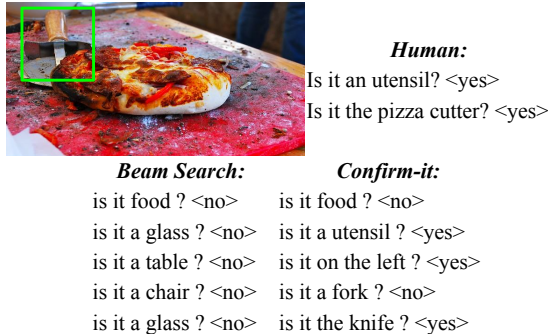is it a glass ? <no>       is it the knife ? <yes>

Figure 2: Through re-ranking, dialogues become more effective and more natural. The target object is highlighted in green.

When the model undergoes the re-ranking phase, *Confirm-it* accuracy has an increase of +4.35% with respect to what it achieves when it outputs the question selected by the plain beam search, and an increase of +4.8% against greedy search. Note that, instead, randomly re-ranking the set of questions lowers the performance. This result shows that confirmation-driven strategies help generate more effective dialogues. Interestingly, our re-ranking method does not require additional training compared to the SL paradigm.

**More Effective Dialogues**    To verify whether the improvement of *Confirm-it* is really due to the generation of more effective dialogues to solve the guessing task, we asked human subjects to guess the target given a dialogue. We sampled 500 games from the GuessWhat?! test set containing less than 7 candidate objects. Each participant played 150 games equally divided among dialogues generated by the model with the plain beam search, with our re-ranking strategy, and by the GuessWhat?! human players (taken from the original test set). We made sure no participant played the same game more than once. In total, 10 English proficient volunteers within our organization joined the experiment. As we can see from Table 2, human annota-

tors reach an accuracy of 70.8% in identifying the target object when receiving dialogues generated by beam search and 77% with *Confirm-it*, suggesting that the re-ranking phase let the model generate more effective dialogues. The accuracy that the annotators achieve when playing the game with dialogues extracted from the original GuessWhat?! test set (and thus generated by human players) is much higher (96%).

**More Natural Dialogues**    Figure 2 reports a sample game that illustrates the difference between a dialogue generated by human players, one generated by the plain beam search, and one by our re-ranking algorithm. The dialogue generated by beam search contains a repetition ("is it a glass?"), it asks about entities not present in the image ("chair" and "glasses") and it ends with a non-conclusive negatively answered question. These features contribute to making the dialogues sound unnatural. We check whether the re-ranking phase helps our model to get closer to human dialogues with respect to these features. To this end, we compute the percentage of games with repeated questions and with the last turn containing a positively answered question. Moreover, we employ CHAIR-s to measure the percentage of hallucinated entities in a sequence, originally proposed in Rohrbach et al. (2018) for image captioning and recently applied also to GuessWhat?! (Testoni and Bernardi, 2021b). CHAIR-S is defined as the number of dialogues with at last one hallucinated entity divided by the total number of dialogues. As we can see from Table 2, dialogues generated by *Confirm-it* contain fewer games with at least one repeated question compared to the beam search setting (-8.14%), fewer games with hallucinated entities (-2.61%)[3], and more games with the last turn con-

---

[3]Rohrbach et al. (2018) propose another variant of the CHAIR metric called CHAIR-i (per-instance), defined as the number of hallucinated objects in a sequence divided by the

taining a positively answered question (71.87% vs. 76.68%). The reduced number of hallucinations is a direct consequence of the *Confirm-it* strategy: following up on a single object through the dialogue, the model is less likely to engage in spurious exchanges on irrelevant objects. Though this strategy continuously looks for confirmations, it is worth noting that it does not increase the number of repetitions, which instead are significantly reduced. This is an interesting property emerging from the interplay between the internal Oracle and the re-ranking strategy, which suggests that asking the very same question more than once in a dialogue does not increase the model's confidence in its hypothesis.

**Qualitative Analysis of the Strategy**  We also evaluated the strategy followed by *Confirm-it* by looking at the model's decisions throughout the dialogue. Interestingly, the model does not select only questions for which it expects a positive answer, though they are the majority (67%). See the SM, for a game in which the re-ranking promoted a question answered negatively by the internal Oracle. Moreover, though the model looks for confirmations, it properly updates its beliefs when disconfirmed: when the model receives from its interlocutor an answer different from the one it expects (based on its internal Oracle), in 70% of the cases the Guesser changes the probabilities over the candidates accordingly, i.e., it assigns the highest probability to a new candidate object. Finally, the use of a human-like strategy does not imply having learned to simply mimic human dialogues from the training set: the re-ranking shows an absolute increase of +12% in the number of pairs of consecutive questions not seen during training (see Table 2).

## 6 Discussion and Conclusion

In this paper, we propose *Confirm-it*, a multimodal conversational model based on a decoding strategy inspired by cognitive studies of human behavior. We show that, through the proposed beam search re-reanking algorithm, our model generates dialogues that are more effective (based on task-accuracy) and more natural (based on the dialogues features discussed above). We believe further improvement could be obtained by increasing the performance of every single module. Moreover, the structure of the generated dialogues remains to be analysed,

---

total number of objects mentioned. CHAIR-i results: 18.28 (Beam Search), 15.02 (*Confirm-it*), 4.11 (Human Dialogues).

and we agree with van der Lee et al. (2021) that a proper evaluation should involve humans. In future work, our method can be easily extended to other task-oriented dialogue tasks which involve a conversational agent as far as it has a module that generates questions and a module that performs a classification task. Depending on the task at hand, different ways to take intermediate probabilities into account can be designed, but the core idea of the method would not change.

## References

Shubham Agarwal, Marc Dymetman, and Éric Gaussier. 2018. Char2char generation with reranking for the E2E NLG challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 451–456, Tilburg University, The Netherlands. Association for Computational Linguistics.

Jonathan Baron. 2000. *Thinking and deciding*. Cambridge University Press, New York.

Frédéric Blain, Pranava Swaroop Madhyastha, and Lucia Specia. 2017. Exploring hypotheses spaces in neural machine translation. *Asia-Pacific Association for Machine Translation (AAMT), editor, Machine Translation Summit XVI. Nagoya, Japan.*

Sebastian Borgeaud and Guy Emerson. 2020. Leveraging sentence similarity in natural language generation: Improving beam search using range voting. In *Proceedings of the Fourth Workshop on Neural Generation and Translation, NGT@ACL 2020, Online, July 5-10, 2020*, pages 97–109. Association for Computational Linguistics.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. GuessWhat?! Visual object discovery through multi-modal dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

Ondřej Dušek and Filip Jurčíček. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages

45–51, Berlin, Germany. Association for Computational Linguistics.

Albert Gatt, Roger PG van Gompel, Kees van Deemter, and Emiel Krahmer. 2013. Are we bayesian referring expression generators. In *Proceedings of the Workshop on Production of Referring Expressions: Bridging the Gap between Computational and Cognitive Approaches to Reference (PRE-CogSci'13).*, Berlin.

James Hargreaves, Andreas Vlachos, and Guy Emerson. 2021. Incremental beam manipulation for natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2563–2574. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Tamara Nicol Medina, Jesse Snedeker, John C Trueswell, and Lila R Gleitman. 2011. How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, 108(22):9014–9019.

Vishvak Murahari, Prithvijit Chattopadhyay, Dhruv Batra, Devi Parikh, and Abhishek Das. 2019. Improving generative visual dialog by answering diverse questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1449–1454, Hong Kong, China. Association for Computational Linguistics.

Wei Pang and Xiaojie Wang. 2020. Visual dialogue state tracking for question generation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11831–11838. AAAI Press.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Azzurra Ruggeri and Tania Lombrozo. 2015. Children adapt their questions to achieve efficient search. *Cognition*, 143:203–216.

Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. 2019. Beyond task success: A closer look at jointly learning to see, ask, and Guess-What. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2578–2587.

Pushkar Shukla, Carlos Elmadjian, Richika Sharan, Vivek Kulkarni, Matthew Turk, and William Yang Wang. 2019. What should I ask? using conversationally informative rewards for goal-oriented visual dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6442–6451, Florence, Italy. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pages 3104–3112.

Alberto Testoni and Raffaella Bernardi. 2021a. The interplay of task success and dialogue quality: An in-depth evaluation in task-oriented visual dialogues. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2071–2082. Association for Computational Linguistics.

Alberto Testoni and Raffaella Bernardi. 2021b. "I've seen things you people wouldn't believe": Hallucinating entities in GuessWhat?! In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 101–111, Online. Association for Computational Linguistics.

John C Trueswell, Tamara Nicol Medina, Alon Hafri, and Lila R Gleitman. 2013. Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology*, 66(1):126–156.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel Krahmer. 2021. Human evaluation of

automatically generated text: Current trends and best practice guidelines. *Comput. Speech Lang.*, 67:101151.

Peter C. Wason. 1960. "on the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3):129–140.

Sina Zarrieß and David Schlangen. 2018. Decoding strategies for neural referring expression generation. In *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands, November 5-8, 2018*, pages 503–512. Association for Computational Linguistics.

Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, Jianfeng Lu, and Anton van den Hengel. 2018. Goal-oriented visual question generation via intermediate rewards. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part V*, volume 11209 of *Lecture Notes in Computer Science*, pages 189–204. Springer.

## A  Supplementary Material

Section 4 of the paper describes the *Confirm-it* model and Figure 3 shows its architecture. In section 5 (Qualitative analysis of the strategy), we highlight that *Confirm-it* does not select only questions for which it expects a positive answer, as shown in Figure 5. In this case, given the dialogue history $H$, the model's hypothesis $c_h$ (the candidate that receives the highest probability according to the Guesser module), and a set of questions $q_1, q_2, q_3$ ordered by their probability according to beam search, the question that helps the most (answered by the internal Oracle taking $c_h$ as the target) is $q_2$. Figure 6 illustrates a step-by-step example of how *Confirm-it* works.
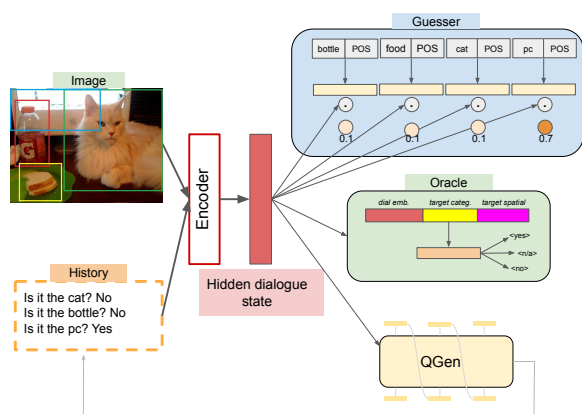


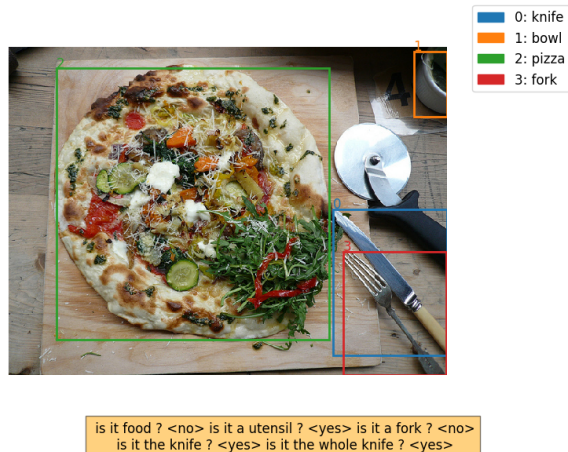Figure 3: Model architecture of *Confirm-it*.



Figure 4: Annotation carried out by human participants.

**GuessWhat?! Dataset Details.** The Guess-What?! dataset contains 155K English dialogues about approximately 66K different images from MSCOCO. The answers' distribution is: 52.2% No, 45.6% Yes, and 2.2% N/A. The training set contains 108K dialogues and the validation and test sets 23K each. Dialogues contain on average 5.2 question-answer pairs. The vocabulary is built up of those words appearing at least 3 times in the training set, resulting in around 4900 words. Each game has at least 3 and at most 20 candidate objects. We train the model using human dialogues, selecting only the games on which humans have succeed in finding the target object in at most 10 turns (total number of dialogues used after this pre-processing step: 90K in training and around 18K both in validation and testing). The dataset of human dialogues is available at https://guesswhat.ai/download.
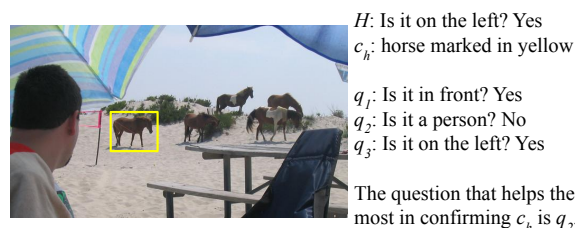


Figure 5: Given an image and the dialogue history $H$, *Confirm-it* assigns the highest probability to $c_h$ (marked in yellow). Beam search generates three questions for the follow-up turn (ordered by their probability): thanks to its internal Oracle, the model answers each of these questions by taking $c_h$ as the target. *Confirm-it* selects $q_2$ (which receives a negative answer according to the internal Oracle) as the question that helps the most in confirming $c_h$.

***Confirm-it* Details.** Regarding the *Confirm-it* model, we used a beam size of 3 in the paper and let the model generate dialogues of 5 turns. We also tried different values for the beam size, obtaining similar results. For the model hyperparameters and training procedure, we follow Shekhar et al. (2019). We trained and tested *Confirm-it* on an NVIDIA TITAN V GPU (12 GB). We used Pytorch 1.0.1 (https://pytorch.org/). *Confirm-it* has 21411226 parameters. The average runtime is 15 minutes per epoch during training and 8 minutes during inference. The maximum number of training epochs is 100. We select the best model by looking at the performance on the GuessWhat?! validation set. The validation accuracy of *Confirm-it* is 51.49 (47.28 with beam search). For the CHAIR metric (Rohrbach et al., 2018), we used the code in: https://github.com/LisaAnne/Hallucination.

**Human Annotation Evaluation.** Figure 4 shows the annotation schema used by human participants in our study, as described in Section 5. The participants in this study are English proficient volunteers within our organization. Each participant is instructed on the guessing task by playing some trial games. The participant is admitted to the annotation only if he/she shows a clear understanding of the task. Given an image, a dialogue and a set of candidate objects with colour-matching boxes, participants express their guess by typing the number corresponding to the box of the selected candidate. Dialogues generated by human annotators from the GuessWhat?! test set, by Confirm-it and by beam search without re-ranking were randomly presented.

**1) Given the dialogue history *H*, compute the candidate objects' probability distribution using the Guesser module and formulate a conjecture about the target (candidate A).**

*H*: <start> is it living? Yes
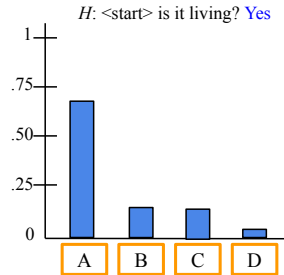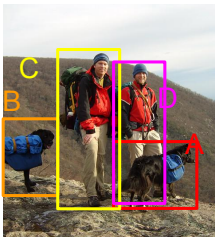
**2) Assume beam size=3. Generate three candidate questions for the follow-up turn starting from *H*. Question *a* is assigned the highest probability according to the beam search algorithm.**

a) is it in the middle?
b) is it a dog?
c) is it a person?

**3) Let the internal Oracle answer each of the three questions as if the candidate with the highest probability at step 1 is the target (answers marked in red). Independently append each QA pair to *H* and compute the candidates' probabilities.**

*H* + is it in the middle? No
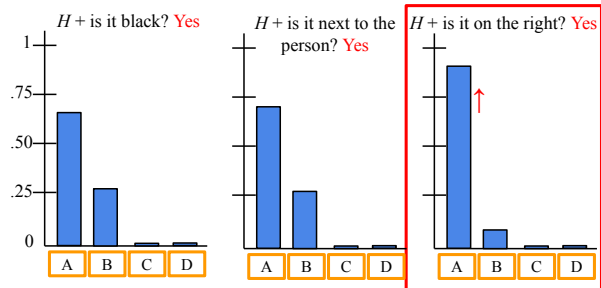*H* + is it a dog? Yes
*H* + is it a person? No
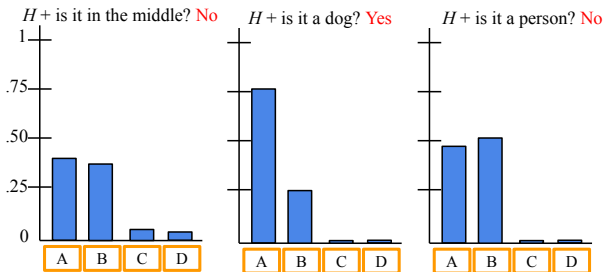
**4) Select the question that, is answered as the model's expectation, confirms the initial conjecture about the target. In this case, 'is it a dog?' increases the most the probability on candidate A if answered positively.**

*H* + is it in the middle? No
*H* + is it a dog? Yes
*H* + is it a person? No
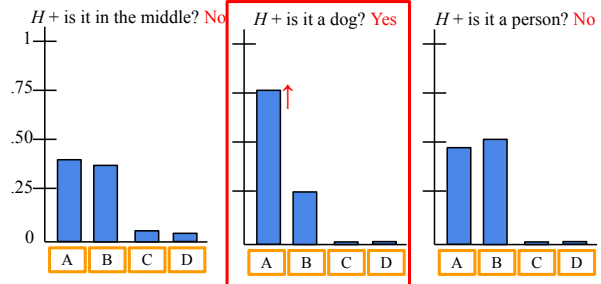
**5) Let the external Oracle who is aware of the real target object answer the previously selected question (answer marked in blue). In this case, the Oracle answers 'yes', in line with model's expectation.**

*H*: <start> is it living? Yes Is it a dog? Yes
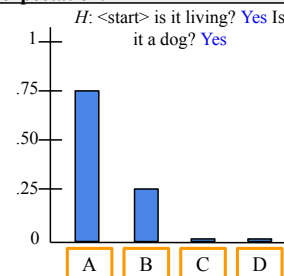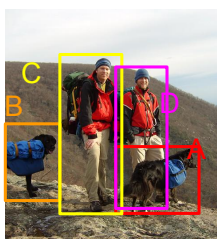
**6) Repeat the steps from 2 to 4 for the follow-up turn.**

*H* + is it black? Yes
*H* + is it next to the person? Yes
*H* + is it on the right? Yes

**7) Let the external Oracle answer the previously selected question. In this case, the Oracle answers 'no', contradicting the model's conjecture about the target.**

*H*: <start> is it living? Yes Is it a dog? Yes is it on the right? No
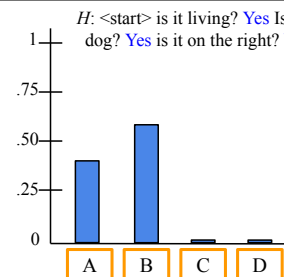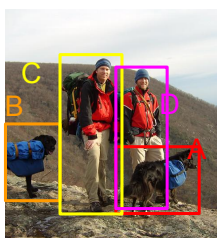
**8) Formulate a new conjecture (candidate B). Repeat the steps from 2 to 5 for the follow-up turns, until the maximum dialogue length is reached.**

**9) At the end of the dialogue, make a guess about the target. Note that *Confirm-it* often selects questions that do not have the highest probability according to the beam search algorithm.**

Figure 6: Step-by-step illustration of how *Confirm-it* works.