Are Current Decoding Strategies Capable of Facing the Challenges of Visual Dialogue?

Amit Kumar Chaudhary
CIMeC, University of Trento
amitkumar.chaudhar@unitn.it

Alex J. Lucassen
CIMeC, University of Trento
alex.lucassen@unitn.it

Ioanna Tsani
CIMeC, University of Trento
ioanna.tsani@unitn.it

Alberto Testoni
DISI, University of Trento
alberto.testoni@unitn.it

Abstract

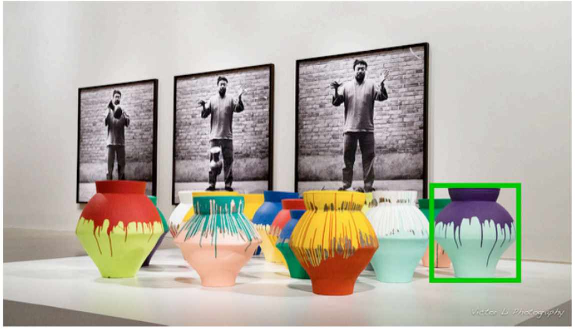
Decoding strategies play a crucial role in natural language generation systems. They are usually designed and evaluated in open-ended text-only tasks, and it is not clear how different strategies handle the numerous challenges that goal-oriented multimodal systems face (such as grounding and informativeness). To answer this question, we compare a wide variety of different decoding strategies and hyper-parameter configurations in a Visual Dialogue referential game. Although none of them successfully balance lexical richness, accuracy in the task, and visual grounding, our in-depth analysis allows us to highlight the strengths and weaknesses of each decoding strategy. We believe our findings and suggestions may serve as a starting point for designing more effective decoding algorithms that handle the challenges of Visual Dialogue tasks.

1 Introduction

The last few years have witnessed remarkable progress in developing efficient generative language models. The choice of the decoding strategy plays a crucial role in the quality of the output (see Zarrieß et al. (2021) for an exhaustive overview). It should be noted that decoding strategies are usually designed for and evaluated in text-only settings. The most-used decoding strategies can be grouped into two main classes. On the one hand, decoding strategies that aim to generate text that maximizes likelihood (like greedy and beam search) are shown to generate generic, repetitive, and *degenerate* output. Zhang et al. (2021) refer to this phenomenon as *the likelihood trap*, and provide evidence that these strategies lead to sub-optimal sequences. On the other hand, stochastic strategies like pure sampling, top-k sampling, and nucleus sampling (Holtzman et al., 2020) increase the variability of generated texts by taking random samples from the model. However, this comes at the cost of generating words that are not semantically appropriate for the context in which they appear. Recently, Meister et al. (2022) used an information-theoretic framework to propose a new decoding algorithm (typical decoding), which samples tokens with an information content close to their conditional entropy. Typical decoding shows promising results in human evaluation experiments but, given its recent release, it is not clear yet how general this approach is.

Multimodal vision & language systems have recently received a lot of attention from the research community, but a thorough analysis of different decoding strategies in these systems has not been carried out. Thus, the question arises of whether the above-mentioned decoding strategies can handle the challenges of multimodal systems. i.e., generate text that not only takes into account lexical variability, but also grounding in the visual modality. Moreover, in goal-oriented tasks, the informativeness of the generated text plays a crucial role as well. To address these research questions, in this paper we take a referential visual dialogue task, GuessWhat?! (De Vries et al., 2017), where two players (a Questioner and an Oracle) interact so that the Questioner identifies the secret object assigned to the Oracle among the ones appearing in an image (see Figure 1 for an example). Apart from well-known issues, such as repetitions in the output, this task poses specific challenges for evaluating decoding techniques compared to previous work. On the one hand, the generated output has to be coherent with the visual input upon which the conversation takes place. As highlighted by Rohrbach et al. (2018); Testoni and Bernardi (2021b), multimodal generative models often generate *hallucinated* entities, i.e., tokens that refer to entities that do not appear in the image upon which the conversation takes place. On the other hand, the questions must be informative, i.e., they must help the Questioner to incrementally identify the target object.

We show that the choice of the decoding strat-



Questioner	Oracle
Is it a vase?	Yes
Is it partially visible?	No
Is it in the left corner?	No
Is it the turquoise and purple one?	Yes

Figure 1: Example of a GuessWhat game from De Vries et al. (2017)

egy and its hyper-parameter configuration heavily affects the quality of the generated output. Our results highlight the specific strengths and weaknesses of decoding strategies that aim at generating sequences with the highest probability vs. strategies that randomly sample words. We find that none of the decoding strategies currently available is able to balance task accuracy and linguistic quality of the output. However, we also show which strategies perform better at important challenges, such as incremental dialogue history, human evaluation, hallucination rate, and lexical diversity. We believe our work may serve as a starting point for designing decoding strategies that take into account all the challenges involved in Visual Dialogue tasks.

2 Task & Dataset

GuessWhat?! (De Vries et al., 2017) is a simple object identification game in English where two participants see a real-world image from MSCOCO (Lin et al., 2014) containing multiple objects. One player (the Oracle) is secretly assigned one object in the image (the target) and the other player (the Questioner) has to guess it by asking a series of binary yes-no questions to the Oracle. The task is considered to be successful if the Questioner identifies the target. The dataset for this task was collected from human players via Amazon Mechanical Turk. The authors collected 150K dialogues with an average of 5.3 binary questions per dialogue. Figure 1 shows an example of a GuessWhat game from the dataset.

3 Model and Decoding Strategies

We use the model and pre-trained checkpoints of the Questioner agent made available by Testoni and Bernardi (2021c) for the GuessWhat?! task. This model is based on the GDSE architecture (Shekhar et al., 2019). It uses a ResNet-152 network (He et al., 2016) to encode the images and an LSTM network to encode the dialogue history. A multimodal shared representation is generated and then used to train both the question generator (which generates a follow-up question given the dialogue history) and the Guesser module (which selects the target object among a list of candidates at the end of the dialogue) in a joint multi-task learning fashion. Testoni and Bernardi (2021c) added an internal Oracle module to the GDSE architecture, which guides a cognitively-inspired beam search reranking strategy (*Confirm-it*) at inference time: this strategy promotes the generation of questions that aim at confirming the model's intermediate conjectures about the target. In our work, at inference time the Questioner agent always interacts with the baseline Oracle agent proposed in De Vries et al. (2017).

We analyse the effect of a large number of decoding strategies as well as hyper-parameter configuration for each strategy: as highlighted by Zhang et al. (2021), it is crucial to evaluate different hyperparameter configurations when comparing multiple decoding strategies. Among the ones that maximize the likelihood of the sequence, we consider plain **beam search** (with a beam size of 3) and **greedy search**. We also consider **Confirm-it**, the cognitively-inspired beam search re-ranking strategy proposed in Testoni and Bernardi (2021c) for promoting the generation of questions that aim at confirming the model's intermediate conjectures about the target. This strategy re-ranks the set of candidate questions from beam search and selects the one that helps the most in confirming the model's hypothesis about the target. As for stochastic strategies, we analyse **pure sampling**, **top-k sampling** (with different k values), and **nucleus sampling** (with different p values), a strategy proposed in Holtzman et al. (2020) which selects the highest probability tokens whose cumulative probability mass exceeds a given threshold p. We also consider **typical decoding** (with different τ values), a recently proposed strategy (Meister et al., 2022) based on an information-theoretic framework. We refer to the respective papers for additional details on decoding strategies. We let the model generate 5 questions[1] at test time and average the results over five random seeds.

4 Metrics

We are interested in evaluating different decoding strategies against a set of metrics that reflect the complexity of the different skills required to successfully solve multimodal referential games.

Linguistic Quality: We compute the percentage of games with at least one repeated question, the overall number of unique words used by the model and, in line with the observations in Testoni and Bernardi (2021a), the number of *rare words* generated by the model, defined as those words that appear fewer than 20 times in the training set.

Visual Grounding: To quantify the rate of object hallucination in the generated dialogues, we compute the CHAIR metric (Rohrbach et al., 2018; Testoni and Bernardi, 2021b). This metric, originally proposed for image captioning, detects hallucination by checking each object mentioned in a generated image caption against the ground-truth MSCOCO objects for that image. The metric consists of two distinct variants: CHAIR-i, or per-instance variant (number of hallucinated objects divided by the total number of objects mentioned in each dialogue), and CHAIR-s, or per-sentence variant (number of dialogues with at least one hallucination divided by the total number of dialogues).[2]

Informativeness: To study the informativeness of the generated questions, we report the raw accuracy of the model in guessing the target object after each dialogue turn and at the end of the dialogue. A game is considered successful if the model identifies the target object assigned to the Oracle. Similarly, we also report the accuracy of human annotators when guessing the target by reading machine-generated dialogues.

5 Results

5.1 Quantitative Results

Table 1 shows the performance of different decoding strategies against accuracy and dialogue quality, as described by the metrics in Section 4. [3] Confirming it is by far the best decoding strategy in terms of accuracy and hallucination rate. However, it uses a restricted vocabulary compared to other strategies. A similar issue is observed for greedy and beam search. We find nucleus sampling (with a p-value of 0.3, much lower than the one used by the authors in Holtzman et al. (2020)) to effectively increase the lexical variety compared to beam search, without damaging accuracy and hallucination rate. Typical decoding, top-k and pure sampling, instead, clearly decrease repetitions and increase the vocabulary richness by generating tokens that are not related to the source input, as indicated by the high hallucination rate. It thus looks like there exists a trade-off between informativeness / visual grounding and linguistic quality.

5.2 Effect of Hyper-Parameter Choice

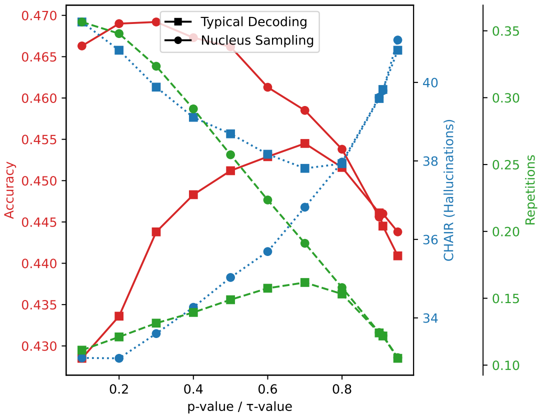


Figure 2: Different hyper-parameter values and their effect on the accuracy, hallucinations, and repetitions in typical decoding and nucleus sampling.

We study the effect of hyper-parameter configurations in stochastic strategies. Specifically, we try various p-values for nucleus sampling and τ-values for typical decoding.[4] As shown in Figure 2, both typical and nucleus sampling peak in accuracy with the parameter configurations that also lead to the most repetitions and fewest hallucinations. Conversely, both strategies show the lowest accuracy with the highest hallucination rate. These results confirm the detrimental effect of hallucinations on the performance of the model. It is interesting to note the robustness of typical decoding in generating few repetitions regardless of the τ value. In line with the findings in Zhang et al. (2021), this analysis confirms the importance of hyper-parameter

[1] Except for the accuracy per turn metric in Section 5.3, where the dialogues consisted of 10 questions.

[2] Testoni and Bernardi (2021b) first adapted the CHAIR metric for Visual Dialogue. However, the authors did not investigate the effect of different decoding strategies.

[3] Here we only report the best-performing configuration for each decoding strategy (see SM for all configurations).

[4] Results for top-k are in SM.

	Accuracy (%) ↑	CHAIR-i ↓	CHAIR-s ↓	% games with repetitions ↓	Vocabulary Size ↑	Rare Words ↑
Confirm-it	**51.39**	**15.09**	**28.48**	30.33	858	34
Beam Search (beam size = 3)	**47.05**	18.33	**31.08**	38.49	731	27
Nucleus Sampling (p = 0.3)	**46.92**	**17.96**	33.60	32.35	1016	78
Greedy Search	46.58	**17.75**	**32.97**	35.63	834	46
Typical Decoding (τ = 0.7)	45.45	21.84	37.81	**16.18**	**1703**	**247**
Top-k Sampling (k = 5)	45.10	22.84	37.71	**14.93**	**1462**	**171**
Pure Sampling	43.13	26.55	43.23	**8.32**	**2609**	**793**

Table 1: Comparison between decoding strategies and their best-performing (in terms of accuracy) hyper-parameters. The decoding strategies are sorted by accuracy.

configurations and the peculiar trade-off between informativeness, repetitions, and visual grounding: so far it has not been possible to find a single configuration that optimizes all three at the same time.

5.3 Per-turn Accuracy

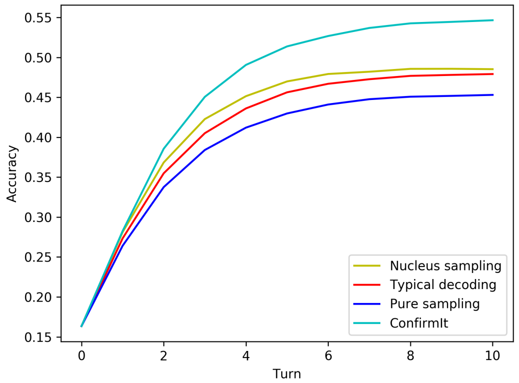


Figure 3: The accuracy per dialogue turn for four different decoding strategies for dialogues of length 10.

One crucial ability in GuessWhat?! is asking informative questions that incrementally help in identifying the target: for this reason, we check the accuracy of the model after each new question is asked. Figure 3 shows accuracy per dialogue turn for a set of representative strategies: Nucleus sampling (p=0.3), Typical Decoding (τ=0.7), Confirm-it, and pure sampling. To get a broader picture, we let the model generate 10 questions in this setting. Confirm-it stands out by showing the largest incremental increase of accuracy throughout the dialogue, indicating that it generates more effective follow-up questions. Pure sampling, on the other hand, seems to suffer from the very beginning of the dialogue and its accuracy stabilizes soon. It is worth noting that the accuracy of typical decoding gets closer to that of nucleus sampling towards the end of the dialogue, with the latter leveling off sooner. We conjecture that Confirm-it outperforms other techniques because it takes into account the probability of the Guesser at inference time, so it is guided to generate questions that change these probabilities and thus avoid generic questions.

	Human Accuracy (%) ↑
Confirm-it	**72.5**
Typical Sampling (τ = 0.7)	68.0
Nucleus Sampling (p = 0.3)	67.5
Pure Sampling	59.5

Table 2: Human Guess Accuracy based on dialogue generated from different decoding strategies.

5.4 Human Evaluation

We asked 8 human annotators to guess the target object in a sample of GuessWhat?! games when reading dialogues generated by our model with different decoding strategies. Each participant annotated 100 games (25 per strategy) and the decoding strategy was not revealed during the annotation. As shown in Table 2, humans reach the highest accuracy when reading dialogues generated by Confirm-it, followed by typical decoding and nucleus sampling, while pure sampling falls behind. These results, which do not mirror the accuracy result in Table 1, allow us to disentangle the weaknesses of the Guesser (i.e., the classification module that predicts the target) from the actual informativeness of the dialogues. Compared to the model, human annotators seem to better exploit the lexical richness of typical decoding and nucleus sampling. We refer to the SM for additional information about the annotation procedure, in line with the best-practice guidelines in van der Lee et al. (2021).

6 Related Work

In the field of multimodal NLG, Zarrieß and Schlangen (2018) propose trainable decoding for referring expression generation. The authors propose a two-stage optimization set-up where a small network processes the RNN's hidden state before passing it to the decoder, using BLEU score as a reward for the decoder. We did not analyse this approach in our paper because we focus only on decoding strategies that do not require any change in the architecture or training of the model. We leave for future work an analysis of trainable decoding approaches. Inspired by the findings in Holtzman et al. (2020), Massarelli et al. (2020) propose a hybrid decoding strategy for open-ended text generation which combines the non-repetitive nature of sampling strategies with the consistency of likelihood-based approaches. The authors show that their approach generated less repetitive and more verifiable text. The design of hybrid decoding strategies for multimodal tasks is out of the scope of this paper, but is an interesting subject to pursue in future work.

7 Discussion and Conclusion

Decoding algorithms are a key component of natural language generation systems. They are usually designed for and evaluated in text-only tasks. We believe multimodal (vision & language) and goal-oriented tasks pose unique and under-studied challenges to current decoding strategies. In this paper, we ran an in-depth analysis of several decoding strategies (and their hyper-parameter configurations) for a model playing a referential visual dialogue game. We found that decoding algorithms that lead to the highest accuracy in the task and the lowest hallucination rate, at the same time generate highly repetitive text and use a restricted vocabulary. Our analyses reveal the crucial role of hyper-parameter configuration in stochastic strategies, an issue that poses several questions about the trade-off between lexical variety, hallucination rate, and task accuracy. While nucleus sampling partially balances the above-mentioned issues, human annotators seem to better exploit the richness of the dialogues generated by typical decoding. Finally, our results demonstrate that a beam search re-ranking algorithm (Confirm-it) generates more effective follow-up questions throughout the dialogue turns. We believe that taking into account the model's intermediate predictions about the referent, like Confirm-it does, represents a promising direction that should be applied also to stochastic strategies in future work, aiming at preserving their lexical richness while reducing hallucinations.

Our results demonstrate that none of the decoding strategies currently at disposal effectively take into account both task accuracy and dialogue quality at the same time. We also highlight peculiar features of each strategy that may guide future research with the goal of designing decoding strategies that properly confront the crucial challenges of multimodal goal-oriented dialogues.

Acknowledgements

We kindly acknowledge the support of the Department of Information Science and Engineering (DISI) for the arrangement of the GPUs used in our research. We would like to thank Professor Raffaella Bernardi for the suggestions and feedback. The first three authors are enrolled in the Erasmus Mundus European Masters Program in Language and Communication Technologies. The first author is supported by the Erasmus+ Programme. The third author of this scientific paper was supported by the Onassis Foundation - Scholarship ID: F ZR 060-1/2021-2022. We thank the Master's students of the Grounded Language Processing course at the University of Trento for the feedback provided at an early stage of this project.

References

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! Visual object discovery through multi-modal dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4466–4475. IEEE Computer Society.

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.

Luca Massarelli, Fabio Petroni, Aleksandra Piktus, Myle Ott, Tim Rocktäschel, Vassilis Plachouras, Fabrizio Silvestri, and Sebastian Riedel. 2020. How decoding strategies affect the verifiability of generated text. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 223–235, Online. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022. Typical decoding for natural language generation. *CoRR*, abs/2202.00666.

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.

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 GuessWhat. 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.

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.

Alberto Testoni and Raffaella Bernardi. 2021c. Looking for confirmations: An effective and human-like visual dialogue strategy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9330–9338, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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. *Computer Speech & Language*, 67:101151.

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

Sina Zarrieß, Henrik Voigt, and Simeon Schüz. 2021. Decoding methods in neural language generation: A survey. *Information*, 12(9):355.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.

A Supplementary Material

A.1 Effect of Hyper-parameters

Figures 4, 5, and 6 illustrate how hyper-parameter choice affects the accuracy, the hallucinations, and the repetitions. Top-k sampling (Figure 4) shows decreased accuracy and repetitions, and increased hallucinations, as the k-value gets higher. The same general pattern can be observed with the gradual increase of the p-value in nucleus sampling (Figure 6). On the other hand, typical decoding accuracy peaks at $\tau = 0.7$ (Figure 5). This is also the point at which the repetitions are at their highest and the hallucinations are at their lowest. Both very high and very low τ-values cause lower accuracy, fewer repetitions, and an increase of hallucinations.

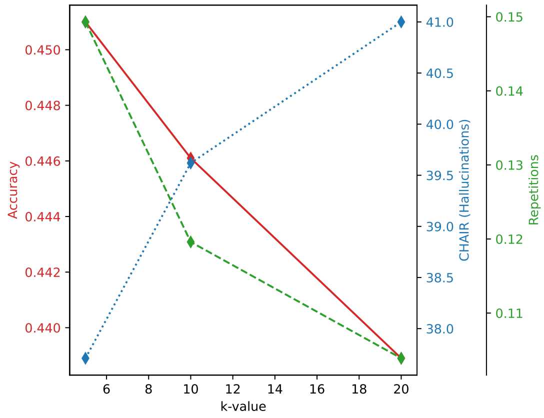


Figure 4: Hyper-parameter choices' effect on the accuracy, hallucinations, and repetitions in top-k sampling.

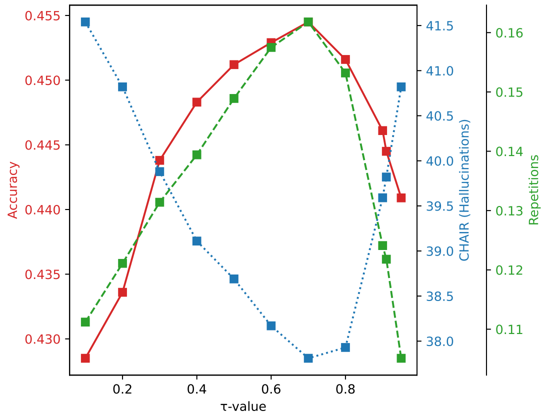


Figure 5: Hyper-parameter choices' effect on the accuracy, hallucinations, and repetitions in typical decoding.

A.2 Experiments

Table 3 presents our results in detail for all the parameter configurations we considered. We have computed accuracy percentage, CHAIR-i, CHAIR-s, percentage of games with repeated questions, vocabulary size and number of rare words for each decoding method and its respective hyper-parameter configurations. These results are sorted by decreasing accuracy. The 3 best results of each metric are in bold.

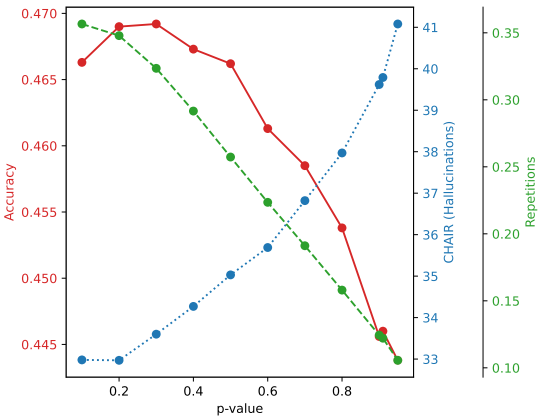
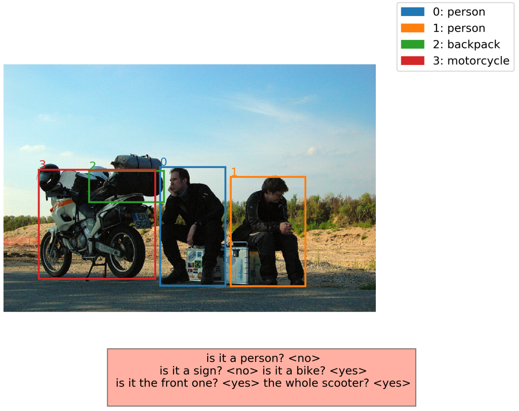


Figure 6: Hyper-parameter choices' effect on the accuracy, hallucinations, and repetitions in nucleus sampling.

A.3 Human Annotation Details



is it a person? <no>
is it a sign? <no> is it a bike? <yes>
is it the front one? <yes> the whole scooter? <yes>

Figure 7: Example of the games displayed to the participants for the annotation task. Participants had to select one target object among the list of candidate objects on the right. The machine-generated dialogue is in the red box.

The annotation was done by 8 human annotators on a sample of GuessWhat?! games. They were recruited within our organization on a voluntary basis and they did not receive any payment for the annotation. Written informed consent was obtained from all the participants. Participants were 4 males and 4 females with high educational level and from

	% Accuracy ↑		CHAIR-i ↓		CHAIR-s ↓		% games with repetitions ↓		Vocabulary Size ↑		Rare Words ↑	
	mean	std	mean	std	mean	std	mean	std	mean	std	mean	std
CI	**51.39**	0.00	**15.09**	0.00	**28.48**	0.00	30.33	0.00	858	0.0	34	0.0
BS (beam = 3)	**47.05**	0.00	18.33	0.00	**31.08**	0.00	38.49	0.00	731	0.0	27	0.0
NS ($p = 0.3$)	**46.92**	0.17	17.96	0.09	33.60	0.14	32.35	0.21	1016	8.3	78	5.5
NS ($p = 0.2$)	46.90	0.11	**17.65**	0.06	**34.79**	0.12	46.41	0.18	919	7.1	59	4.5
NS ($p = 0.4$)	46.73	0.27	18.38	0.17	34.27	0.28	29.16	0.25	1097	11.9	88	6.4
NS ($p = 0.1$)	46.63	0.02	17.76	0.01	32.98	0.01	35.66	0.06	839	1.6	46	1.9
NS ($p = 0.5$)	46.62	0.17	19.10	0.09	35.03	0.10	25.73	0.26	1192	18.0	103	3.4
GS	46.58	0.00	**17.75**	0.00	**32.97**	0.00	35.63	0.00	834	0.0	46	0.0
NS ($p = 0.6$)	46.13	0.38	20.04	0.15	35.69	0.35	22.35	0.26	1303	12.7	126	9.4
NS ($p = 0.7$)	45.85	0.19	21.14	0.11	36.82	0.31	19.11	0.26	1451	9.0	162	9.5
TD ($\tau = 0.7$)	45.45	0.32	21.84	0.15	37.81	0.23	16.18	0.29	1703	13.0	247	12.6
NS (p = 0.8)	45.38	0.14	22.20	0.16	37.97	0.30	15.80	0.19	1643	23.0	219	12.9
TD ($\tau = 0.6$)	45.29	0.28	22.08	0.16	38.17	0.30	15.75	0.17	1723	21.9	248	18.3
TD ($\tau = 0.8$)	45.16	0.18	22.21	0.20	37.93	0.29	15.32	0.28	1712	10.8	244	13.6
TD ($\tau = 0.5$)	45.12	0.15	22.60	0.17	38.69	0.36	14.89	0.22	1745	7.3	262	8.7
Top-k ($k = 5$)	45.10	0.27	22.84	0.21	37.71	0.26	14.93	0.10	1462	12.6	171	5.2
TD ($\tau = 0.4$)	44.83	0.17	23.11	0.13	39.11	0.44	13.94	0.24	1755	19.0	265	12.2
TD ($\tau = 0.9$)	44.61	0.16	23.74	0.18	39.59	0.19	12.41	0.25	1919	13.5	334	9.1
Top-k ($k = 10$)	44.61	0.24	24.03	0.29	39.62	0.26	11.96	0.16	1692	13.8	235	10.5
NS ($p = 0.91$)	44.60	0.15	23.92	0.13	39.79	0.23	12.21	0.13	1948	22.3	342	14.0
NS ($p = 0.9$)	44.56	0.23	23.82	0.07	39.62	0.17	12.44	0.10	1912	20.2	332	13.6
TD ($\tau = 0.91$)	44.45	0.27	23.92	0.14	39.82	0.31	12.18	0.19	1945	11.7	345	16.1
TD ($\tau = 0.3$)	44.38	0.31	24.07	0.21	39.88	0.22	13.14	0.23	1791	14.8	278	13.7
NS ($p = 0.95$)	44.38	0.12	24.93	0.15	41.08	0.24	10.56	0.05	**2129**	11.3	**438**	11.4
TD ($\tau = 0.95$)	44.09	0.21	24.83	0.24	40.82	0.28	**10.51**	0.20	**2117**	18.9	**435**	17.1
Top-k ($k = 20$)	43.89	0.10	25.12	0.30	41.00	0.39	**10.39**	0.17	1879	23.1	305	17.5
TD ($\tau = 0.2$)	43.36	0.20	25.19	0.16	40.82	0.34	12.11	0.09	1815	21.5	287	10.7
PS	43.13	0.28	26.55	0.25	43.23	0.36	**8.32**	0.17	**2609**	9.3	**793**	11.4
TD ($\tau = 0.1$)	42.85	0.15	26.25	0.14	41.54	0.09	11.12	0.11	1825	18.4	286	13.5

Table 3: Comparison between decoding strategies and their hyper-parameters (CI = Confirm-it, BS = Beam Search, NS = Nucleus Sampling, GS = Greedy Search, TD = Typical Decoding, Top-k = Top-k Sampling, PS = Pure Sampling).

different ethnic groups. Before the beginning of the annotation task, each annotator was briefed on the GuessWhat?! gameplay and purpose, and was asked to annotate some sample games in order to get familiar with the annotation process. We used the `makesense.ai` online software for image recognition. Each image had a minimum of 3 and a maximum of 6 candidate objects. The annotators could see both the bounding box and the category for each candidate object in the image. They could also see the full dialogue between the Questioner and the Oracle. The annotators then had to pick the object they believed was the right one, based on the information given by the dialogue. Figure 7 provides an example of the games we asked the participants to annotate. Overall, we extracted 200 images from the GuessWhat?! test set and generated one dialogue per decoding strategy per image. In total, we thus generated 800 dialogues. Each human participant annotated 25 images per decoding strategy. To prevent biases, the participants were not given any information about the decoding strategy that each image represented.

A.4 Comparison of Decoding Strategies

Table 4 shows a comparison between generated dialogues of various decoding strategies (see Figure 8 for the image the dialogues refer to).

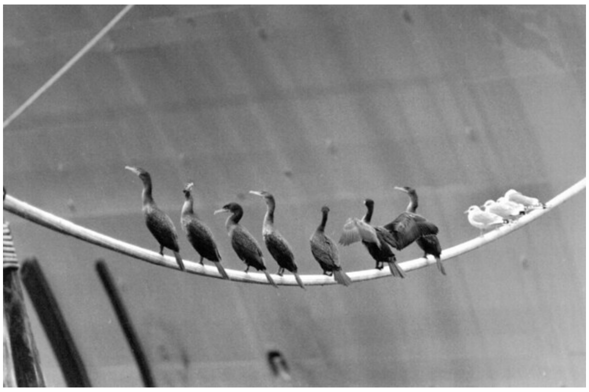


Figure 8: The image for game 147344.

Turn	Speaker	Confirm-it	Nucleus Sampling	Typical Decoding	Pure Sampling
0	Questioner	is it a bird?	is it a bird?	is it a bird?	is it a bird?
	Oracle	yes	yes	yes	yes
1	Questioner	is it on the left side?	is it on the left side?	is it on the air?	is it close?
	Oracle	yes	yes	yes	yes
2	Questioner	is it on the left side of the group?	is it one of the first birds?	is it 1 of the 3 on the right?	is it squatting?
	Oracle	no	yes	no	yes
3	Questioner	is it in the middle?	is it the first one from the left?	is it one of the three birds that are next to each other?	is it one of the three all the way to the right?
	Oracle	yes	no	yes	yes
4	Questioner	is it one of the three on the right?	is it the second one from the left?	is it further left?	is it 1st?
	Oracle	no	no	yes	no

Table 4: Comparison between dialogues generated by various decoding strategies for the image reported in Figure 8.