ALOHa: A New Measure for Hallucination in Captioning Models

Anonymous ACL submission

Abstract

Despite recent advances in multimodal pre-training for visual description, state-of-the-art models still produce captions containing errors, such as hallucinating objects that are not present in a scene. The existing prominent metric for object hallucination, CHAIR, is limited to a fixed set of MS COCO objects and synonyms. In this work, we propose a modernized open-vocabulary metric, ALOHa, which leverages large language models (LLMs) to measure object hallucinations. Specifically, we use an LLM to extract groundable objects from a candidate caption, measure their semantic similarity to reference objects from captions and/or object detections, and use Hungarian matching to produce a final hallucination score. We show that ALOHa correctly identifies 13.6% more hallucinated objects than CHAIR on HAT, a new gold-standard subset of MS COCO Captions annotated for hallucinations, and 30.8% more on nocaps, where objects extend beyond MS COCO categories.

1 Introduction and Background

In recent years, vision-language models have demonstrated remarkable performance. Unfortunately, even state-of-the-art models for visual description still generate captions with object hallucinations – objects or entities that are present in the caption yet are not explicitly supported by visual evidence in the image (Dai et al., 2023). In order to reduce the occurrence of object hallucinations in vision-language models, it is helpful to understand and quantify the problem through *reliable*, *localizable*, and *generalizable* measures of object hallucination. *Reliable* measures are capable of correctly indicating if a given caption contains an object hallucination. *Localizable* measures are capable of indicating which object in a particular caption is hallucinated. *Generalizable* measures are capable of evaluating captions from a wide range of input datasets, across a wide range of object and entity categories.

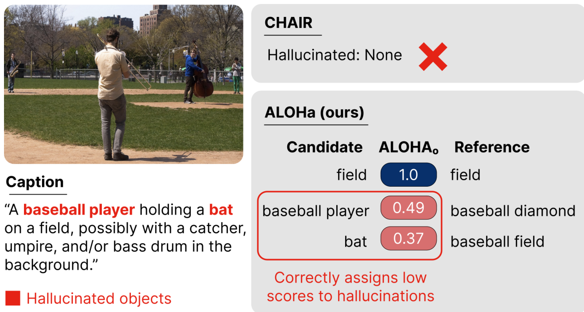

Figure 1: (Top) The SOTA prior object hallucination metric, CHAIR, is limited to MS COCO objects, and fails to detect the hallucinations in this image caption while ALOHa (ours, bottom) correctly assigns low similarity scores to the hallucinations "baseball player" and "bat". ALOHa does not penalize the caption for "catcher", "umpire", and "bass drum", as the caption indicates uncertainty of their presence.

Recent works that measure object hallucinations in generated text generally fall into two categories: measures that find hallucinations by explicitly matching from a set of objects, and measures that compute distances between latent image and/or text embeddings, indicating a hallucination if the embeddings are too distant. In the first category, CHAIR (Rohrbach et al., 2018) is a measure that explicitly extracts objects from candidate sentences using simple string matching against MS COCO classes and a small set of synonyms. It compares these extracted objects against the ground truth detections, and objects extracted from the ground truth reference captions. CHAIR is both reliable, as string matching on a fixed set of objects is accurate and consistent, and localizable, as individual non-matching strings are identified. However, as seen in Figure 1, CHAIR is not generalizable, as it can only handle a fixed set of predetermined objects. Other uni-modal measures in this category include those for abstractive summarization (Durmus et al., 2020; Kryscinski et al., 2020; Maynez et al., 2020; Son et al., 2022; Sridhar and Visser, 2022; Yuan et al., 2021), dialogue (Huang et al., 2022; Shuster et al., 2021), and structured knowledge (Dhingra et al., 2019). These often generalize poorly to vision-language tasks as they require grounding the generated text into inputs of the same modality.

In the second category, CLIPScore (Hessel et al., 2021) employs CLIP (Radford et al., 2021) embeddings to assess image-text matches. While it is generalizable and reliable, it lacks localization as it does not pinpoint incorrect spans of text. CLIPBERTS (Wan and Bansal, 2022) and Ref-CLIPScore (an extension of CLIPScore accounting for reference captions) face similar limitations.

POPE (Li et al., 2023) evaluates vision-language models' likelihood to hallucinate objects with machine-generated queries consisting of samples extracted from both reference object detections and nonexistent objects. The POPE approach, while a useful object hallucination score for model comparison, addresses a fundamentally different problem from that which we investigate here – it measures how often *models* hallucinate rather than localizes and detects hallucinations within *a single caption*.

Inspired by recent successes using LLMs for evaluation in language-only tasks (Zhang et al., 2020; Yuan et al., 2021; Bubeck et al., 2023; Chiang et al., 2023; Zheng et al., 2023), we introduce Assessment with Language models for Object Hallucination (ALOHa), a modernized measure for object hallucination detection that is *reliable*, *localizable*, and *generalizable*. ALOHa extends the reliability and localization of CHAIR to new input domains by leveraging in-context learning of LLMs combined with semantically-rich text embeddings for object parsing and matching (Figure 1).

For a given image caption, we generate two measures: **ALOHa$_\mathbf{o}$**, a numeric score for each object rating the degree to which that object is a hallucination, and **ALOHa**, an aggregated score rating the degree to which the whole caption contains a hallucination. We demonstrate the performance of ALOHa on a new gold-standard dataset of image hallucinations, HAT, and show that ALOHa is more accurate than CLIPScore at detecting object hallucinations, and more accurate than CHAIR at correctly localizing those hallucinations. We conclude by demonstrating that ALOHa remains reliable and localizable when generalizing to out of domain data.

2 ALOHa: Reliable, Localizable, and Generalizable Hallucination Detection

ALOHa produces numeric scores rating the degree of hallucination for each object in a candidate caption as well as an overall caption score, given a set of ground-truth reference captions and predicted (or ground truth) image object detections. ALOHa consists of three stages (Figure 2). (1) Objects are extracted from the image, reference set, and candidate caption using a combination of an object detector and LLM. (2) We filter the object sets and compute semantic representations of each object. (3) We compute a maximum-similarity linear assignment between candidate and reference objects. The scores from each of the pairs in the linear assignment, which we call ALOHa$_\mathrm{o}$, measure the degree of hallucination for each of the candidate objects. The minimum similarity in this linear assignment (the ALOHa score) measures the degree of hallucination of the caption.

(1) Extracting objects from candidates, references, and images: Parsing visually grounded objects in a caption in an open-domain context is a surprisingly difficult task. CHAIR (Rohrbach et al., 2018) relies on a fixed set of MS COCO objects and synonyms, requiring considerable effort to extend to other datasets, and sometimes failing at ambiguous parses (such as mistaking the adjective "orange" for a noun). SPICE (Anderson et al., 2016) relies on standard grammar-based object parsing, which can have similar issues, as purely text-based methods fall short at identifying which nouns are *visual* – for instance, avoiding "picture" and "background" in Figure 2. Captions may also indicate uncertainty around object presence, such as "a bowl or plate", or "a dog biting something, possibly a Frisbee." We aim to handle such uncertain objects to avoid unfair hallucination penalties.

With the understanding that open-domain parsing is the primary factor in CHAIR's lack of generalization, we leverage the capability of zero-shot in-context learning in large language models. Following Brown et al. (2020), we use an LLM (ChatGPT, OpenAI (2022)) along with the prompt given in Appendix A to turn the parsing task into a language completion task easily solvable by an LLM. We encourage the LLM to extract visual objects in the scene, consisting primarily of noun phrases (including any attributes, such as "big dog" and "purple shirt"), from the candidate and reference captions. We run the LLM against the candidate caption to produce the unfiltered object set $\mathcal{C}$, and again for the corresponding reference captions to produce object set $\mathcal{R}$. To extract objects from the image context, similar to CHAIR, we

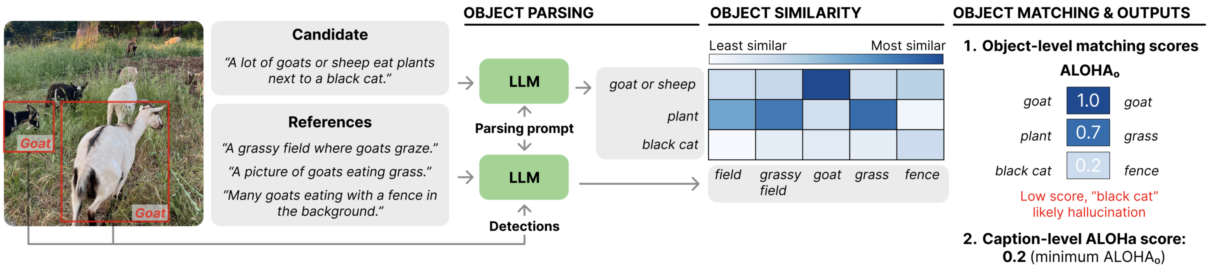

Figure 2: Overview of ALOHa. We prompt an LLM to extract visually-grounded nouns from a candidate machine-generated description and a set of references. We consider uncertain language (e.g., *"goat or sheep"*), add reference objects with and without modifiers (e.g., both *"field"* and *"grassy field"*), and avoid non-visual nouns (e.g., *"picture"* and *"background"*). Then, we compute a maximum-similarity linear assignment between candidate and reference object sets, the weights of which form the ALOHa$_o$. Matched pairs with low ALOHa$_o$ are likely hallucinations (e.g., *"black cat"*, ALOHa$_o = 0.2$). We additionally output the minimum ALOHa$_o$ as a caption-level ALOHa score.

augment the set of reference objects with objects detected directly from the image using DETR (Carion et al., 2020) fine-tuned on MS COCO.

(2) Object filtering: We further refine candidate ($\mathcal{C}$) and reference ($\mathcal{R}$) object sets to better reflect specific challenges of object hallucination detection. Ideally, hallucination measures should penalize specificity when candidate attributes are not supported by references (e.g., if "purple shirt" $\in \mathcal{C}$, yet "white shirt" $\in \mathcal{R}$), but should not penalize generality (e.g., "shirt" $\in \mathcal{C}$, yet "white shirt" $\in \mathcal{R}$). Thus, we use spaCy (Honnibal et al., 2020a) to augment $\mathcal{R}$ with the root nouns from each *reference* noun phrase, but leave the candidates unchanged.

Beyond specificity, captions may also express uncertainty about the presence of objects in an image. For conjunctions (e.g., "fork or knife"), we aim to avoid unfair penalties if at least one of the objects is grounded. ALOHa considers all combinations of selecting a single object from each conjunction, denoted as $\mathcal{C}_{\{1...M\}}$ and $\mathcal{R}_{\{1...N\}}$ (e.g., "fork" $\in \mathcal{R}_0$ and "knife" $\in \mathcal{R}_1$). Additionally, we prompt the LLM to indicate uncertain grounding by including "possibly" after the object (e.g., "there may be a Frisbee" becomes "Frisbee (possibly)") and we remove uncertain objects from $\mathcal{C}_i$ to avoid penalties while maintaining them in $\mathcal{R}_j$ for maximum coverage of more general objects.

(3) Object Matching: Once we have extracted and parsed the candidate and reference object sets, we aim to measure the degree of hallucination for each candidate object. While we could match objects based on string alone (resulting in a binary decision), as does CHAIR, often it is useful to understand a continuous scale of hallucination – e.g., for a reference object "dog", hallucinating "wolf" should be penalized less than "potato." To capture this scale of semantic similarity, for each object text o, we generate $o_{\text{emb}} = \phi(o) \in \mathbb{R}^K$, where ϕ is a semantic text embedding model. In our work, we use S-BERT (Reimers and Gurevych, 2019). We then compute a similarity score for each pair of objects (usually the cosine similarity, see Appendix B.3). For each $(\mathcal{C}_i, \mathcal{R}_j)$ pair, we store these scores in a similarity matrix $\mathcal{S}_{i,j} \in [0,1]^{|\mathcal{C}_i| \times |\mathcal{R}_j|}$. We then use the Hungarian method (Kuhn, 1955) to find an optimal maximum-similarity assignment $\mathcal{M}_{i,j}$ between candidate and reference sets of objects.

To determine the ALOHa$_o$ score for each object, we take the maximum score across all possible parsings, giving the candidate caption the benefit of the doubt, for an object $c \in \mathcal{C}_i$

$$\text{ALOHa}_o(c) = \max_{i,j} w_{c_i,j} \in \mathcal{M}_{i,j} \tag{1}$$

While $0 \leq \text{ALOHa}_o \leq 1$ indicates the degree of hallucination for each object, we also want to indicate if an entire caption contains a hallucination. We thus define:

$$\text{ALOHa} = \min_{c \in \mathcal{C}} \text{ALOHa}_o(c) \tag{2}$$

We choose the minimum as the presence of *any* hallucinated object indicates that the full caption is a hallucination, and even several correct detections should not compensate for a hallucination.

3 Evaluation & Discussion

HAT: To promote the development of high-quality methods for hallucination detection, we collect and release HAT (HAllucination Test), a dataset of labeled hallucinations in captions. HAT consists of 490 samples (90 validation and 400 test) labeled by in-domain experts for hallucination on both a word level and caption level (See Appendix D). Measures are evaluated on two metrics: Average Precision (AP) and Localization Accuracy (LA). The AP

Method	LA	AP
Baseline (Majority Vote)	-	33.75
CHAIRs	6.70	36.85
CLIPScore	-	40.10
RefCLIPScore	-	48.40
ALOHa (No Detections)	19.55	48.40
ALOHa (Oracle Detections)	19.55	47.86
ALOHa (DETR Detections)*	20.30	48.62
ALOHa (Oracle+DETR Detections)	**21.05**	**48.78**

Table 1: Test set performance for binary hallucination detection on HAT. LA: Localization Accuracy. AP: Average Precision. * indicates the version of ALOHa used throughout this paper, unless noted otherwise. Oracle detection are human-generated reference detections.

of the method measures reliability, and is defined as how well the measure identifies captions with hallucinations. For CHAIR, decisions are binary, so AP = accuracy. For ALOHa, AP is the weighted mean of precisions across all thresholds. The LA, measured on samples containing hallucinations in HAT, measures localization and is defined as the accuracy of correctly indicating *which* of the specific objects were hallucinated. For CHAIR, a hallucination is correctly localized when at least one detected string mismatch is a hallucination, and for ALOHa when the minimum ALOHa$_o$ score corresponds to a hallucinated object.

ALOHa's performance on HAT is shown in Table 1. On AP, ALOHa with DETR detections outperforms both CHAIR and CLIPScore by 11.8% and 8.5% respectively. RefCLIPScore attains a similar AP; however, is not localizable. ALOHa achieves more than twice the LA on HAT CHAIR, a particularly challenging task as HAT includes non-object hallucinations, such as incorrect verbs or relations (see Figure A7). Table 1 further ablates the choice of image detections, and indicates that ALOHa is robust to missing detections.

FOIL object hallucinations: To indicate generalizability we evaluate our method on two machine-generated object hallucination datasets. FOIL (Shekhar et al., 2017) contains MS COCO images, where objects are randomly replaced with similar ones (e.g., "bus" and "car"), and nocaps-FOIL, a similar dataset that we construct on the nocaps dataset (Agrawal et al., 2019) for novel object captioning beyond MS COCO (see Appendix D.1). While both methods are strong on the FOIL dataset, CHAIR fails to transfer to the nocaps-FOIL dataset, as the object set becomes out of scope. CHAIR achieves an AP of only 58.33 (only slightly better than chance) and LA of 14.42, compared to

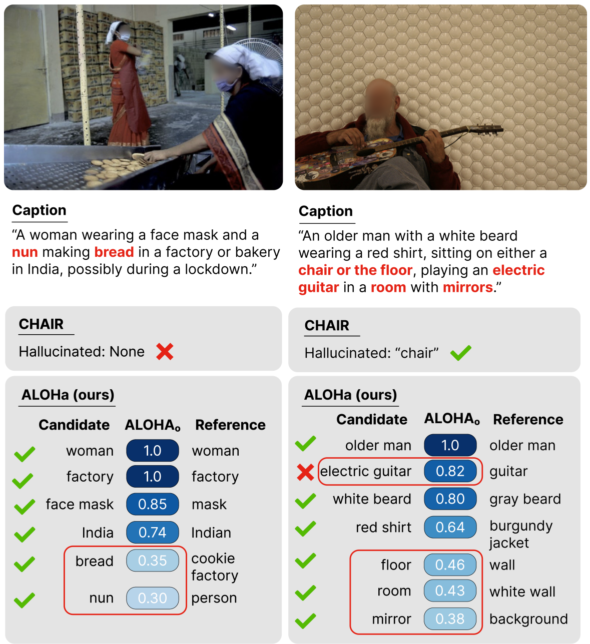

Figure 3: Qualitative Flickr30k examples. (Left) ALOHa correctly assigns low scores to the hallucinated "nun" and "bread", whereas CHAIR does not detect any hallucinations. (Right) Although ALOHa assigns high similarity between the hallucinated "electric guitar" and reference "(acoustic) guitar", it assigns low scores to the other 3 hallucinations. CHAIR detects only the hallucination "chair", missing the others.

ALOHa's AP of 69.52, and LA of 45.17 (213% relative improvement). See Appendix C.2 for details.

Flickr30k: In Figure 3 and Figure A5, we visualize the behavior of CHAIR and ALOHa on several Flickr30k samples (Young et al., 2014), using captions generated by a recent captioning model (Chan et al., 2023) that often produces complex captions with phrases expressing uncertainty.

Additional results: As LLMs can hallucinate themselves (Bubeck et al., 2023), we annotate the parsing error rate on HAT in Table A1 and find that GPT-3.5 introduces extraneous objects in 2.97% of samples. In Appendix B we investigate the choice of LLM, similarity measure, and parsing approach.

4 Conclusion

This paper introduces ALOHa, a scalable LLM-augmented metric for open-vocabulary object hallucination. ALOHa correctly identifies 13.6% more hallucinated objects on HAT and 31% on nocaps-FOIL than CHAIR. ALOHa represents an important modernization of caption hallucination metrics, and detecting complex hallucinations in actions, quantities, and abstract concepts remains an exciting and challenging task for future exploration.

5 Limitations

While ALOHa represents a strong step towards open-domain localized hallucination detection, it comes with several limitations which we discuss in this section.

Non-determinism A primary concern with using large language models for an evaluation measure is the natural nondeterminism that comes with them. While in theory language models sampled at a temperature of zero (as we do in this work) are deterministic, it is well documented that small random fluctuations can still occur (OpenAI, 2023). Beyond random fluctuations, the availability of language models long-term can impact the reproducibility of the measure. In this work, we primarily rely on closed source language models, which can change or become unavailable without notice. In Table A1, we demonstrate that ALOHa still functions with open source models such as Koala (Geng et al., 2023), however the performance is significantly degraded due to the parsing capabilities of the model. With time, and more powerful open source LLMs, this will become less of an issue, however relying on a nondeterministic metric for comparative evaluation can easily become a liability.

Availability of Reference Captions (Reference-Free vs. Reference-Based Measures) One of the primary limitations of the ALOHa evaluation method is the requirement that reference captions are available for the evaluation dataset (an issue shared by CHAIR). Not only must reference captions be available, but they also must sufficiently cover the salient details in the reference image. When the references are impoverished (as can easily happen with a single reference sentence (Chan et al., 2023)) or when there are no references, and ALOHa must rely entirely on detections, the method under-performs more general methods such as CLIPScore which are reference free, and rely on a large pre-training dataset to encode vision and language correspondences. We strongly believe that the area of reference-free localized hallucination detection is an important area of future research; how can we leverage the tools from large vision and language pre-training in a localized way to understand and interpret where hallucinations lie in hallucinated text? That being said, there is also a place for reference-based measures, as reference-based measures focus on what *humans* believe to be salient details in the image, whereas reference-free measures always rely on downstream models which *approximate* what humans believe to be important. This means that reference-based measures can often transfer better to new domains than reference-free measures, which often must be trained/fine-tuned in-domain with human-labeled data to achieve strong performance.

General costs associated with LLMs The use of large language models for any task incurs significant compute, monetary, environmental, and human costs . ALOHa is a significantly slower evaluation measure than methods like CHAIR (however not that much less efficient than CLIPScore), leading to increased power consumption, and cost during evaluation. In addition, the models that we rely on are generally closed source, and represent a non-trivial monetary expenditure (Experiments in this paper, including ablations, testing, and prototyping required approximately $120 USD in API fees). Such factors can be limiting to researchers who wish to evaluate large datasets, however we hope that with the advent of larger open source models, and continued investment in hardware and systems research, the cost will decrease significantly. Beyond compute and financial costs, there are environmental and human costs associated with using large language models for evaluation, see Bender et al. (2021) for a detailed discussion of these factors.

Limited Control of Bias In this work, we do not evaluate the performance of ALOHa on Non-English data, nor do we explicitly control for or measure bias in the creation of HAT (Which is a labeled subset, randomly selected of the MS COCO dataset), or the Nocaps-FOIL dataset (which operates on the same samples as the Nocaps validation dataset). While HAT is a subset of the common MS COCO dataset, we recognize that the creation of such potentially biased datasets has the potential to lead researchers to engineer features and methods which are unintentionally biased against underrepresented groups. We aim to address these shortcomings in the next iteration of HAT, which will not only contain out of domain data for MS COCO trained models, but also aims to better control for bias in the underlying image and caption data. Note that our work, including HAT, is intended for research purposes.

References

Harsh Agrawal, Peter Anderson, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8947–8956. IEEE.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *ArXiv preprint*, abs/2303.12712.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer.

David M Chan, Austin Myers, Sudheendra Vijayanarasimhan, David A Ross, and John Canny. 2023. Ic3: Image captioning by committee consensus. *ArXiv preprint*, abs/2302.01328.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2136–2148, Dubrovnik, Croatia. Association for Computational Linguistics.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. Blog post.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020a. spacy: Industrial-strength natural language processing in python.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020b. spacy: Industrial-strength natural language processing in python, zenodo, 2020.

Sicong Huang, Asli Celikyilmaz, and Haoran Li. 2022. Ed-faith: Evaluating dialogue summarization on faithfulness. *ArXiv preprint*, abs/2211.08464.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023. Gpt-4 technical report.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada. Association for Computational Linguistics.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Seonil Simon Son, Junsoo Park, Jeong-in Hwang, Junghwa Lee, Hyungjong Noh, and Yeonsoo Lee. 2022. Harim+: Evaluating summary quality with hallucination risk: Evaluating summary quality with hallucination risk. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 895–924.

Arvind Krishna Sridhar and Erik Visser. 2022. Improved beam search for hallucination mitigation in abstractive summarization. *ArXiv preprint*, abs/2212.02712.

David Wan and Mohit Bansal. 2022. Evaluating and improving factuality in multimodal abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9632–9648, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5288–5296. IEEE Computer Society.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv preprint*, abs/2306.05685.

Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. 2023. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *ArXiv preprint*, abs/2303.06594.

Appendix

Appendix A describes the prompt of the language model, including the exact language used, the design choices, and the in-context examples.

Appendix B describes several explorations of the hyperparameters, including the language model chosen for ALOHa, the semantic embedding method, and the parsing approach.

Appendix C contains additional detailed results and experimental details for experiments in the paper.

Appendix D describes the datasets that we collected and constructed, including HAT and nocaps-FOIL.

A Prompt

The choice of prompt for a large language model using in-context learning is critical to the performance of the model. Each component of the prompt has some ability to shape the downstream language distribution. In this work, we use the prompt shown in Figure A1. This prompt has several rules, which we discuss here.

Attributes: We ask that the language model include all attributes attached to the object if they are present. By doing so, we can catch hallucinations such as those shown in Figure 3, where "electric guitar" appears in the candidate, but an acoustic guitar is shown in the image. Attributes are handled differently between reference captions and candidate captions. For reference captions, we add both the object with attributes, and the object without attributes to the set, so the candidate is not penalized for being more general. For the candidate, however, we add only the object with attributes, so if the candidate produces attributes, they must match with something in the reference set.

```
You are an assistant that parses
visually present objects from
an image caption. Given an image
caption, you list ALL the objects
visually present in the image or
photo described by the captions.
Strictly abide by the following
rules:
- Include all attributes and
adjectives that describe the object,
if present
- Do not repeat objects
- Do not include objects that
are mentioned but have no visual
presence in the image, such as light,
sound, or emotions
- If the caption is uncertain
about an object, YOU MUST include
'(possibly)' after the object
- If the caption thinks an object can
be one of several things, include
'or' and all the possible objects
- Always give the singular form of
the object, even if the caption uses
the plural form
```

Figure A1: The prompt that we use for parsing objects from both captions and sets of reference captions.

Repeated Objects: In this work, our primary goal is to determine if a particular object is hallucinated, and not focus on the quantity of hallucinations. Thus, we de-duplicate the object set in both the candidate and reference captions, as well as detections coming from the image. By doing this, we focus on if the objects can possibly exist in the image, rather than focus on getting the exact count, which may be incorrect if a candidate caption mentions the same object more than once (and that object is parsed twice).

Intangible Object: In many cases, objects mentioned in the candidate or reference set may be intangible, such as color, light, sound, or emotion. To improve the accuracy of the model, we explicitly suggest that such objects should not be included.

```
Caption: This image shows two pink
roses in a tulip-shaped vase on a
wooden kitchen counter, next to a
microwave and a toaster oven.
Objects:
- pink rose
- tulip-shaped vase
- wooden kitchen counter
- microwave
- toaster oven
```

Figure A2: An example of a single-caption parsing result.

Or/Possibly: Modern captioning methods such as Chat-Captioner (Zhu et al., 2023) and IC3 (Chan et al., 2023) are capable of encoding uncertainty into their approach through the use of words like “possibly" or “maybe". Additionally, they may make judgments that are uncertain such as “an apple or an orange." Existing captioning and hallucination detection measures fail to account for this uncertainty, and match both objects, even though the semantics of the caption suggests that the object is uncertain, or may be one of many objects. To account for this, we encourage the LLM to indicate uncertainty in a fixed way, as well as list multiple alternatives on a single line. We then account for this in our matching method, by giving the candidate the benefit of the doubt, scoring only the best match from an alternative set, and ignoring any uncertainty.

Singularization: While it is possible to singularize objects using rule-based methods, rule-based methods struggle with challenging nouns, and we found that in general, the LLM was better at performing the singularization set of the post-processing before object matching.

A.1 In-Context Examples

In addition to the core prompt text, we provide several contextual samples, which help with in-context learning (Brown et al., 2020). The contextual samples help to align the label space of the model correctly with the target output distribution (Min et al., 2022). An example of such contexts is given in Figure A2 and Figure A3.

B Hyperparameter Exploration

In this section, we explore the choices of hyperparameters for ALOHa including the object parsing, semantic embedding, and language model.

```
Captions:
- Several people riding on a
motorcycle with an umbrella open.
- Couples riding motor cycles
carrying umbrellas and people
sitting at tables.
- A group of people riding scooters
while holding umbrellas.
- Some tables and umbrellas sitting
next to a building.
- Pedestrians and motorcyclists near
an open outdoor market.
Objects:
- person
- couple
- motorcycle
- umbrella
- table
- scooter
- building
- pedestrian
- motorcyclist
- open outdoor market
```

Figure A3: An example of a multi-caption parsing result.

B.1 Object Extraction and Semantic Embedding Methods

In the primary work, we leverage LLMs (OpenAI, 2023) for object extraction, and a BERT-based model (Reimers and Gurevych, 2019) for semantic word embedding. These are, however, not the only choices that can be made for these two components. In Figure A4, we explore the difference in overall performance on HAT's validation set when using different combinations of object extraction and semantic embedding. Namely, we compare LLM-based extraction to the parse-tree-based noun extraction in SpaCy (Honnibal et al., 2020b), and compare SentenceTransformer (BERT-Based model, (Reimers and Gurevych, 2019)) to Word2Vec (Mikolov et al., 2018), GPT-3 (Ada) embedding, and string matching (strings are case-normalized and lemmatized). In general, we found that combining LLMs with the Sentence-Transformer (BERT-Based) model performed better than other methods, and that fuzzy embedding methods often significantly outperform exact string matching when determining hallucination as judged by human raters. This is generally expected: humans have a wide vocabulary that is poorly captured

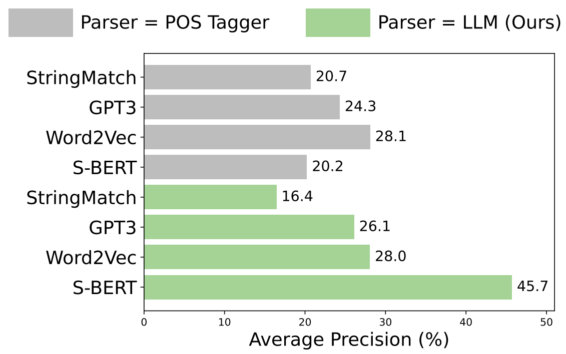

Figure A4: Performance on HAT validation set filtered for hallucinated objects, when comparing embedding methods and object extraction approaches.

by exact string matching. Interestingly, Word2Vec outperforms GPT-3 embeddings. We believe that this is because the GPT-3 embeddings are optimized for sentence-level structures, and may fail to semantically embed single words in a meaningful way.

B.2 Choice of Large Language Model

The choice of the language model is critical to the overall performance of ALOHa- if the language model does not have sufficient zero-shot parsing capability, it will lead to reduced downstream performance. We investigate the performance of the language model in Table A1 on HAT. In these experiments, we measure the average precision (AP) and LA (see Appendix C.1), as well as the "Parsing error rate" (PER), which is the rate of errors made when parsing objects from reference captions on HAT. We calculate PER (Parsing Error Rate) with manual annotation by taking the fraction of objects output by the LLM that did not exist in the caption (in other words, measuring 1-precision of parsed objects). We additionally annotate and compute the recall - the fraction of objects in the caption that are included in the objects parsed by the LLM. This gives a recall for GPT-3.5 of 98.63%. In these experiments, we find that while Koala (Geng et al., 2023) has strong LA performance on HAT, however ChatGPT (GPT-3.5) (OpenAI, 2023) has both the best average precision, and makes the fewest errors, thus we leverage GPT-3.5 for our primary experiments in the main paper.

B.3 Semantic Similarity Measure

In ALOHa, we compute the similarity between objects using the cosine distance between embedding vectors generated using the `all-MiniLM-L6-v2` S-BERT implementation in the Sentence-

Lanugage Model	LA ↑	AP ↑	PER ↓	PRR ↑
GPT-3.5	20.30	**48.62**	**2.97**	**98.63**
Claude (Instant)	20.74	41.48	3.31	-
Koala	**22.22**	38.70	5.07	-

Table A1: Exploration of LLM choice for parsing within ALOHa, on HAT. AP: Average Precision, LA: Localization Accuracy, PER: Parsing Error Rate (%), PRR: Parsing Recall Rate.

Transformers[1] library (Reimers and Gurevych, 2019). While in theory cosine distances should lie in the interval $[-1,1]$, in this library, for optimization stability, models are trained with positive samples having similarity 1, and negative samples having similarity 0. This (unintentionally) induces a model which (by optimization) only produces positive cosine similarity scores. ALOHa can still be adapted to negative similarity: our algorithms for maximal assignment and equations 1 and 2 both support negative values (even though they don't appear in this instantiation of the algorithm).

It is further worth noting that S-BERT is not a word similarity measure, and was instead designed to measure distances between sentences. This means that the underlying metric may not be sufficiently optimized at a word level, however, in Figure Figure A4, we give an ablation on the parsing method and examined the effectiveness of different embedding models for semantic similarity. We found among the explored approaches that S-BERT was most effective and that simple word embedding methods such as GLoVe are insufficient. That being said, we acknowledge that leveraging a model trained specifically for semantic similarity between words would be an exciting and powerful extension to the ALOHa framework. With the development of better word embedding models for semantic similarity, we see greater potential in localizing hallucinations with ALOHa.

C Experimental Details & Additional Experimentation

C.1 Metrics

We employ several measures in the paper, which we describe in detail here.

Average Precision We measure the **Average Precision (AP)** of each hallucination metric to detect sentence-level hallucinations. Specifically,

[1]https://www.sbert.net/

Method	FOIL Overall LA	FOIL Overall AP	nocaps-FOIL In-Domain LA	In-Domain AP	Near-Domain LA	Near-Domain AP	Out-Domain LA	Out-Domain AP	Overall LA	Overall AP
CHAIRs	**79.00**	**92.50**	13.47	57.82	17.55	59.14	12.24	58.06	14.42	58.33
CLIPScore	-	76.44	-	71.81	-	70.17	-	78.73	-	**73.48**
RefCLIPScore	-	80.64	-	**79.63**	-	**78.70**	-	**85.89**	-	**81.31**
ALOHa	40.00	61.35	**47.35**	71.80	**47.30**	66.67	**48.84**	70.91	**45.17**	69.52

Table A2: Breakdown of results by domain on nocaps FOIL. AP: Average Precision. LA: Localization Accuracy. Bold and underlined values represent the best and second-best methods respectively.

we label each sample with **1** if it contains a hallucination and **0** otherwise. We then measure AP between those labels and per-sample hallucination measures. For ALOHa, this is:

$$\text{AP} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\text{label}]\cdot\text{ALOHa}(i) \tag{3}$$

For CHAIR, this is:

$$\text{AP} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}[\text{label}]\cdot\mathbb{I}[\text{CHAIR Prediction}] \tag{4}$$

Localization Accuracy Localization accuracy (LA) measures the fraction of samples where a metric can correctly identify a hallucinated object, among samples that are known to contain hallucinated objects.

$$\text{LA} = \frac{|\{\geq 1 \text{ correctly identified halluc.}\}|}{|\{\geq 1 \text{ halluc.}\}|} \tag{5}$$

A sample receives a LA of 1 if at least one of the predicted hallucinated objects was correct (for CHAIR), or if the object with the minimum matching score was a true hallucination (for ALOHa). We do not measure LA for CLIPScores, as they do cannot provide hallucination scores per object.

C.2 FOIL

Table A2 breaks down the results of ALOHa on the FOIL and nocaps-FOIL dataset. The results illustrate a set of subtle results: while ALOHa under-performs CHAIRs in both AP and LA on the original FOIL dataset, this is primarily due to the construction of the dataset itself. FOIL constructs new samples by replacing string-matched COCO objects with a set of hand selected "foil" objects, objects that are visually distinct from the original object, but are near semantic neighbors. This is a best case scenario for CHAIR, as CHAIR relies on string matching alone, and thus, is easily able to both detect and localized the replaced samples. The inaccuracies in LA and AP come from the synonym set that CHAIR uses for matching, along with parsing errors such as parsing the color "orange" as the object "orange". In much the way that FOIL is the perfect dataset for CHAIR, FOIL perfectly exploits the strengths of ALOHa. Because of the semantic similarity score, we assign less weight to in-domain hallucinations, making it less likely that the replaced FOIL objects will be detected.

When we move to nocaps-FOIL with non-MS COCO data, however, we see starkly contrasting results. ALOHa significantly outperforms CHAIR, as now the object set that was a strength for in-domain FOIL becomes a liability, and CHAIR is unable to detect any hallucinations at all, due to the restricted string matching. RefCLIPScore, while extremely competitive in the hallucination detection task, cannot perform localization, and thus, serves only as a benchmark for the performance of ALOHa on the FOIL and nocaps-FOIL datasets.

D Datasets

In this section, we discuss further the data that we use and go into detail on the dataset collection process for HAT (Appendix D.2) and the nocaps-FOIL dataset (Appendix D.1)

D.1 nocaps-FOIL

The FOIL dataset (Shekhar et al., 2017) is a synthetic hallucination dataset based on samples from the MS-COCO (Xu et al., 2016) dataset. In this dataset, for each candidate-image pair, a "foil" caption is created which swaps one of the objects (in the MS-COCO detection set) in the caption with a different, and closely related neighbor (chosen by hand to closely match, but be visually distinct). While the FOIL dataset provides a useful benchmark for many hallucination detection methods, it is overly biased towards methods optimized for the MS-COCO dataset. To help evaluate methods that are more general, we introduce a new dataset "nocaps-FOIL" based on the nocaps (Agrawal et al., 2019) dataset. The nocaps dataset consists of images from the OpenImages (Kuznetsova et al., 2020) dataset annotated with image captions in a similar style to MS-COCO. nocaps is split into three sets: an in-domain set, where objects in the images are in the MS-COCO object set, near-domain, where the objects in the image are related to those of MS-COCO, and out-of-domain, where objects in the image are not contained in MS-COCO.

To build the nocaps-FOIL dataset, for each image, we generate the baseline caption by removing a single caption from the reference set. We then generate the foil caption as follows. First, we find any words in the baseline caption that are contained in either the openimages class list (there are 600) or a near neighbor in wordnet. We then randomly select one of these classes to replace. Because there are 600 classes, we do not hand-pick the foil classes, and rather, select a near neighbor class based on sentence embeddings from (Reimers and Gurevych, 2019). We find that in practice, the nearest neighbor is often a synonym, thus, to avoid selecting synonyms, we take the 10th furthest sample, which is often a near neighbor, but is visually distinct. We replace this word in the caption, matching case, and then perform a filter for grammatical correctness using the Ginger[2] API. Any captions which are not grammatically correct are filtered. This leaves us with 2500 image/caption/foil pairs, which we use for evaluation in Table A2.

The OpenImages dataset annotations are under a CC BY 4.0 license, and the images are under a CC BY 2.0 license.

D.2 HAT

HAT is based on MS-COCO and aims to be a gold-standard benchmark for the evaluation of hallucination in image captioning methods. While it is relatively small, it is densely annotated by in-domain experts for several types of hallucination including object hallucination, action hallucination, and numeric hallucination among others. HAT consists of 90 validation samples, and 400 test samples, each containing a machine candidate caption generated by one of BLIP (Li et al., 2022), OFA (Wang et al., 2022), IC3 (Chan et al., 2023) or Chat-Captioner (Zhu et al., 2023), and annotations which mark which word in the captions are hallucinated (See Figure A8 for exact instructions given to annotators). An image/caption pair is considered a hallucination if at least one of the words in the caption is hallucinated.

Screenshots of the interface for data collection are given in Figure A8. While initial versions of the dataset were collected using AMT workers, we found that the quality of annotations was not sufficiently high, and thus, trained experts explicitly in hallucination detection, and leveraged expert ratings for the samples in the test dataset.

MS-COCO is under a Creative Commons Attribution 4.0 License.

E Qualitative Examples

We provide additional qualitative examples from the following scenarios:

E.1 Flickr30k Examples

Figure A5 shows several examples on the Flickr-30k dataset Young et al. (2014) with captions generated by IC3 (Chan et al., 2023), a modern image captioning model that often generates longer, more complex captions including uncertain language such as "possibly." We highlight objects with $\text{ALOHa}_o \leq 0.5$ as likely hallucinations. For samples going from left to right:

1. The caption hallucinates the word "mother", as there is no visual evidence that the woman is specifically a mother. CHAIR does not capture this, as "mother" is mapped to a synonym for "person", which it counts as a grounded (non-hallucinated) object. ALOHa matches "mother" to the reference "person", assigning a borderline ALOHa_o of 0.5.

2. The image does not contain a hallucination. CHAIR flags "table" as hallucinated, yet the caption expressed uncertainty with a conjunction: "chair or table." ALOHa successfully parses this conjuction and selects "cloth" with $\text{ALOHa}_o = 1.0$ to the exact reference match.

3. CHAIR does not detect the hallucinated "bridge", which is successfully assigned a low $\text{ALOHa}_o = 0.35$.

[2] https://www.gingersoftware.com/

4. The caption hallucinates the word “father”. In most cases, the specific relationship of “father” is unlikely to be grounded (similar to “mother” in sample 1); yet, in this image, it is even more clear as there are only children present. CHAIR maps “father” as another synonym for “person” and does not consider it a hallucination, whereas “father” has a low $\text{ALOHa}_\text{o} = 0.34$.

E.2 HAT Examples

We present 4 random samples from HAT each for cases without hallucinations (Figure A6) and with hallucinations (Figure A7). Because these examples contain more nuance that we discuss below, we do not indicate binary hallucination decisions as in Appendix E.1.

Starting with Figure A6), samples with captions that were labeled as correct, from left to right:

1. Both CHAIR and ALOHa successfully do not find any hallucinations.

2. CHAIR does not flag any hallucinations. ALOHa assigns a low $\text{ALOHa}_\text{o} = 0.36$ for “sun“, an incorrect parse from the phrase “sunny day”. However, the other objects are successfully matched. Interestingly, ALOHa adds “snowboard” as an object, inferring that the physical item would need to be present given the verb “snowboarding”.

3. CHAIR again does not flag any hallucinations. ALOHa_o for “tall building” is the mid-range 0.59, matched with the reference “building”, indicating a somewhat uncertain attribute. This may be reasonable given the point of view in the image.

4. CHAIR finds no hallucinations. “Cloudy sky” receives a somewhat low $\text{ALOHa}_\text{o} = 0.45$. Although this phrase is accurate given the image, this is a failure case in which the references are incomplete.

Next, we discuss Figure A7, showing samples that were labeled to contain a hallucination. Recall that labels capture *all* types of caption errors, including those other than object hallucinations, to serve as a valuable source for research around general caption correctness. As a result, there exists non-object hallucinations in HAT that are impossible for CHAIR or ALOHa to localize. From left to right:

1. The attribute “tall” is labeled as a hallucination, as the building next to the bus is only one story. Similar to sample 3 in Figure A6, ALOHa_o for “tall building” is somewhat uncertain at 0.59. Other objects are correctly grounded.

2. The object “table” is a hallucinated, misclassified object; e.g., one reference opts for the more general “wooden surface.” However, the reference mentions a “table” that it is placed on, leading CHAIR to avoid considering it as a hallucination. For ALOHa, this example shows one of the 2.97% of cases (Table A1) where ALOHa hallucinates a reference object, “dining table”. The candidate “round wooden table” is matched to it, with an erroneously high ALOHa_o of 0.74.

3. This sample contains a complex error, in which the arrow is not, in fact, “pointing in different directions.” This non-object hallucination is impossible for the object-specific CHAIR and ALOHa to localize correctly. However, it demonstrates ALOHa’s capability to extract more complex attributes such as “red street sign” and “orange detour sign.”

4. The cat’s location “on top of a small chair” is labeled as an error. CHAIR does not flag any hallucinations. ALOHa_o for “small chair” is 0.59, yet both metrics cannot capture the specific relation.

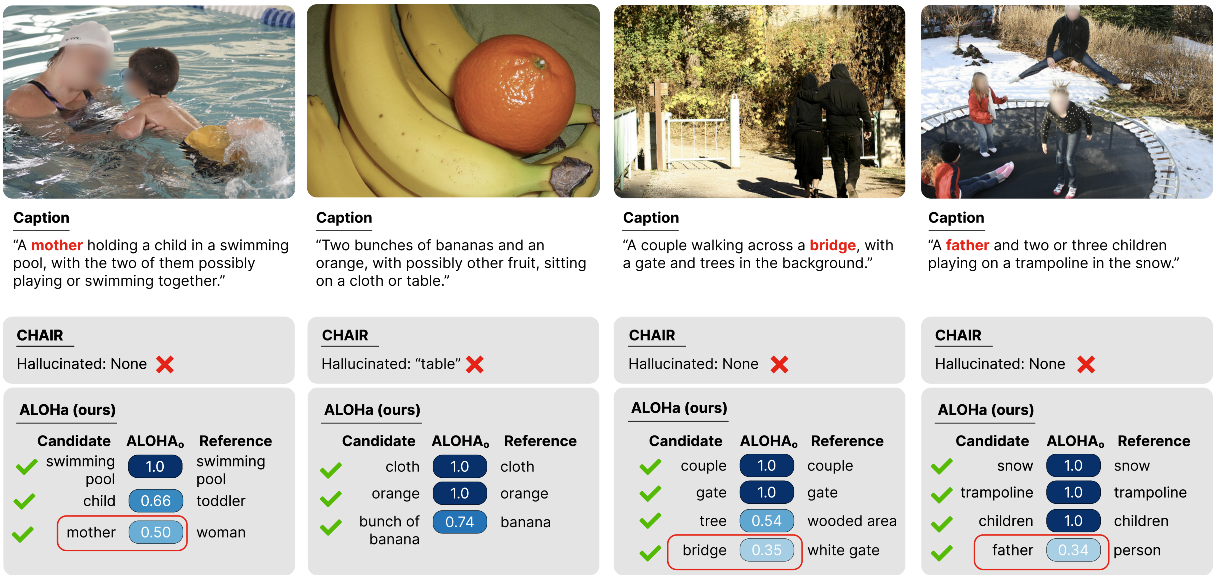

Figure A5: Qualitative samples of ALOHa evaluated on the Flickr-30k dataset, with candidate captions generated by IC3 (Chan et al., 2023). Hallucinated objects in the caption text are red and bolded. See Appendix E.1 for discussion.

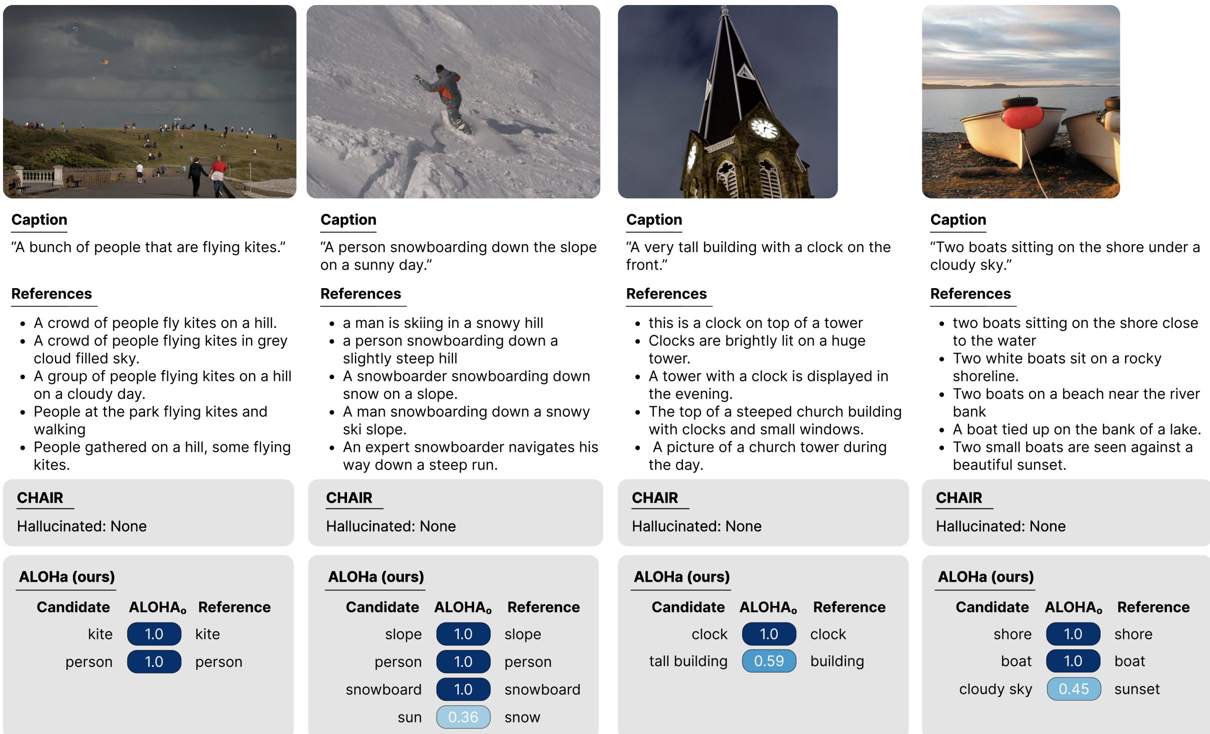

Figure A6: Randomly selected qualitative examples of ALOHa evaluated on the HAT dataset when there is no hallucination in the ground truth. See Appendix E.2 for discussion.

Caption

"A bus driving down a street next to a **tall** building."

References

- A large long bus on a city street.
- A city bus on the street in front of buildings.
- A blue bus traveling down an incline of a busy street.
- A city bus with full side advertisement in front of a building.
- a public transit bus on a city street

CHAIR

Hallucinated: None

ALOHa (ours)

Candidate	ALOHA_o	Reference
bus	1.0	bus
street	1.0	street
tall building	0.59	building

Caption

"A round wooden **table** with a small pizza."

References

- A platter with a baked good on it
- A plain piece of bread resting on a wooden plate.
- A whole cheese pizza sitting on a wood pan on a table.
- a close up of a pizza on a wooden surface on a table
- A white cracker looking pizza is on a cutting board.

CHAIR

Hallucinated: None

ALOHa (ours)

Candidate	ALOHA_o	Reference
round wooden table	0.74	dining table
small pizza	0.69	pizza

Caption

"A street sign with a detour **pointing in different directions**."

References

- An orange detour sign hanging from a metal pole under a cloudy sky.
- Red street sign with black letters sitting on metal post.
- A street pole with an orange detour sign.
- a close up of a street sign with a sky background
- A red detour sign that is on a pole.

CHAIR

Hallucinated: None

ALOHa (ours)

Candidate	ALOHA_o	Reference
street sign	0.83	red street sign
detour	0.59	orange detour sign

Caption

"A cat stands **on top of a small chair**."

References

- A cat perched on top of a dresser. A cat walks along the top of a bedroom dresser.
- a cat sits on a dresser next to a rocking chair
- Black cat standing on a blue dresser next to a chair.
- A cat laying on top of a blue dresser near a chair.

CHAIR

Hallucinated: None

ALOHa (ours)

Candidate	ALOHA_o	Reference
cat	1.0	cat
small chair	0.59	chair

Figure A7: Randomly selected qualitative examples of ALOHa evaluated on the HAT dataset when there is a hallucination in the ground truth. These hallucinations are generally challenging to detect. See Appendix E.2 for discussion.

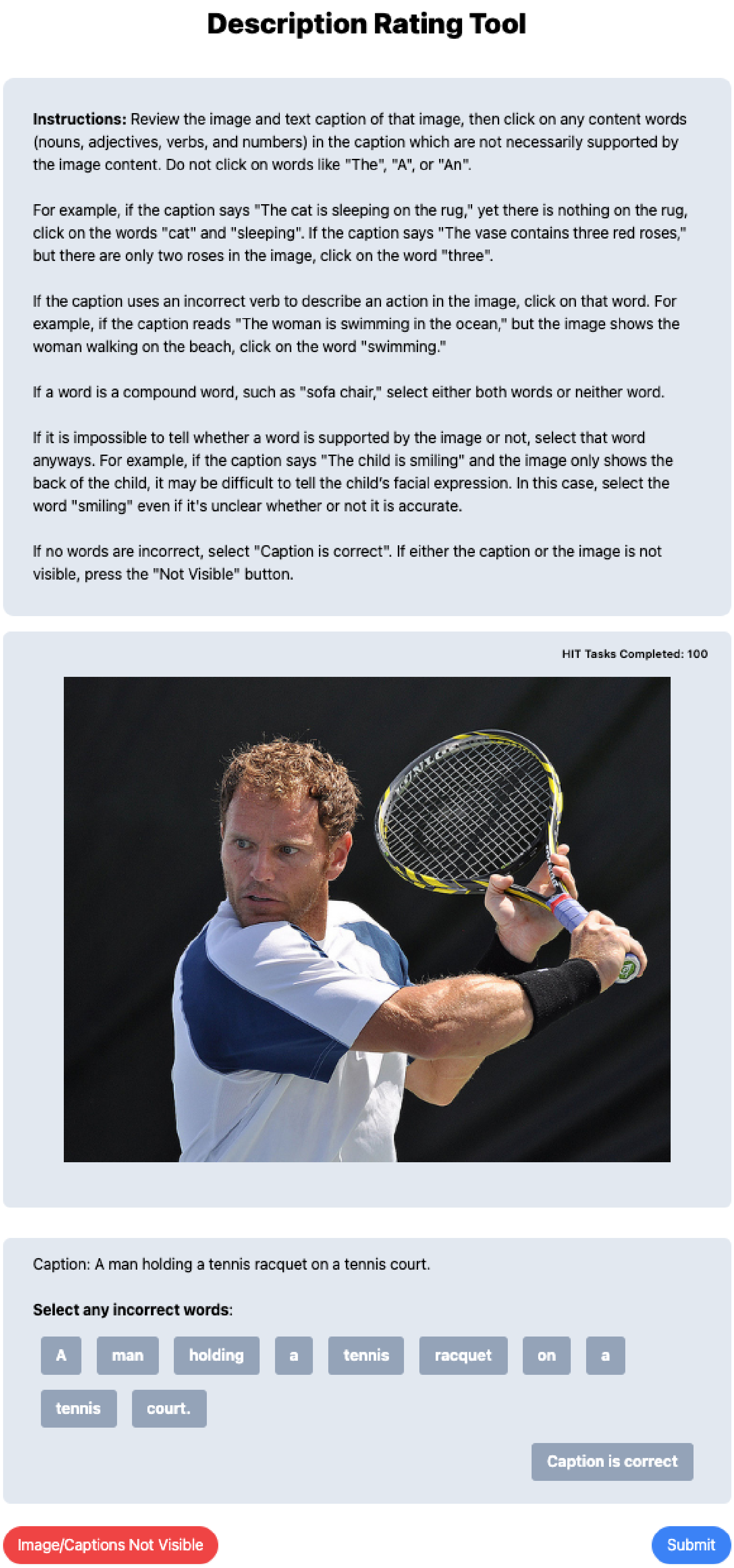

Figure A8: The hallucination dataset collection interface.