Does Object Grounding Really Reduce Hallucination of Large Vision-Language Models?

Anonymous ACL submission

Abstract

Large vision-language models (LVLMs) have recently dramatically pushed the state of the art in image captioning and many image understanding tasks (e.g., visual question answering). Nonetheless, LVLMs still often *hallucinate* and produce captions mentioning concepts that cannot be found in the input image. These hallucinations erode the trustworthiness of LVLMs and are one of the main obstacles to their ubiquitous adoption. Recent work suggests that addition of grounding objectives such as those based on *referring expressions*—explicit alignment between image regions or objects and text descriptions— reduces the amount of LVLM hallucination. Although intuitive, this claim is not empirically justified as the reduction effects have been established, we argue, with flawed evaluation protocols that (i) rely on data (i.e., MSCOCO) that has been extensively used in LVLM training and (ii) measure hallucination via question answering rather than open-ended caption generation. In this work, in contrast, we offer the first systematic analysis of the effect of fine-grained object grounding on LVLM hallucination under an evaluation protocol that more realistically captures LVLM hallucination in open generation. Our extensive experiments reveal that, while grounding leads to more informative captions, it generally does not reduce the proportion of hallucinated content.

1 Introduction

Large Vision-Language Models (LVLMs) have displayed impressive image understanding (Li et al., 2023a; Liu et al., 2023c; Bai et al., 2023; Fini et al., 2023; OpenAI, 2023; Anil et al., 2023, inter alia). Their widespread adoption, however, is hindered by the *object hallucination* problem in which the LVLMs—similar to "general" hallucinations of LLMs (Zhang et al., 2023b)— "invent" objects (or attributes, relations, etc.) not present in the image.

A range of methods have been proposed for this problem like modified decoding strategies (Leng et al., 2023; Huang et al., 2023), post-hoc removal of hallucinations (Yin et al., 2023; Zhou et al., 2023), or reinforcement learning (Sun et al., 2023; Zhao et al., 2023b; Gunjal et al., 2023; Yu et al., 2023). Most approaches, however, either increase inference cost or need expensive additional training and/or data, limiting their ubiquitous applicability.

A recent line of work (Chen et al., 2023b; You et al., 2023; Pramanick et al., 2023) has suggested that including *grounding objectives*—e.g., based on referring expressions (Kazemzadeh et al., 2014) where textual descriptions of image regions have to be grounded to the respective parts of the image— into the LVLM training reduces object hallucination. The claim is intuitive: The region-level objectives are expected to encourage a finer-grained image understanding than 'global' image captioning, the de-facto main training objective of LVLMs, as has been shown for image-text compositionally (Bugliarello et al., 2023), and should discourage the model from generating content it cannot ground in the image. However, despite being intuitive, the empirical support for reduced hallucinations is lacking and mainly stems from evaluation in QA scenarios where the model has to decide if objects are (not) present in an image (Li et al., 2023b); we argue that this evaluation protocol aligns poorly with real-world *free-form* generative applications like image captioning where there is no evidence yet that grounding objectives reduce hallucination.

Contributions. In this work, we perform the first comprehensive analysis of the effects that grounding objectives have on LVLM object hallucination in open-form image captioning, addressing the shortcomings of prior hallucination evaluation protocols. Concretely, we measure the effects of two grounding objectives, added as additional objectives to standard image captioning-based training of LVLMs: (1) *referring data* objective asks the model to generate the bounding box of the region that corresponds to a textual description and vice versa; whereas the (2) *caption grounding* objective demands that the model generates image descriptions with interleaved bounding boxes for mentioned objects. We then compare the extent of hallucination for LVLM variants trained with and without the grounding objectives. To this end, we compare the hallucination measures based on question answering (QA) (Li et al., 2023b) against open-ended free-form metrics (Rohrbach et al., 2018; Jing et al., 2023). Crucially, since (Rohrbach et al., 2018; Li et al., 2023b) rely on MSCOCO (Lin et al., 2014) but MSCOCO is also used for training LVLMs and they are thus a priori less likely to hallucinate on these examples, we extend the evaluation to out-of-distribution datasets. To this end, we propose an alternative method for CHAIR using semantic comparison that addresses the shortcomings of string matching.

Findings. Our experiments confirm that object grounding reduces hallucination in a QA-based protocol; at the same time, in free-form generation, we find no reduction in hallucination: this, we believe, questions the utility of QA-based hallucination evaluation like Li et al. (2023b). Our analysis reveals that *referring data* greatly increases how informative captions are, that is, they contain more descriptive content, but that this also results in increased hallucinations. On the other hand, *caption grounding* leads to shorter captions – this can be seen as a decrease in hallucination, but one that comes at the expense of less informative captions; Neither of the two grounding objectives consistently reduces hallucination. Overall, we find that, while grounding objectives do improve fine-grained image understanding of LVLMs, this does not translate into less hallucination in open caption generation.

2 Grounding Objectives in LVLMs

Grounding objectives seek to align natural language expressions with regions in the image. These objectives either take image regions as input, commonly in the form of a bounding box, predicting corresponding natural language expressions or produce such regions as output. A range of LVLMs have been proposed in recent times that include grounding tasks in their training mix alongside other objectives like captioning and visual question answering (Liu et al., 2023b; Bai et al., 2023; Wang et al., 2023b); other models have been designed specifically for expression grounding and trained with grounding objectives only (Chen et al., 2023b; You et al., 2023; Pramanick et al., 2023; Zhang et al., 2023a; Peng et al., 2023; Chen et al., 2023a; Zhao et al., 2023a).

Objectives. Our investigation focuses on the two arguably most popular grounding objectives, commonly included in LVLMs training: referring expressions (Kazemzadeh et al., 2014) and grounded captions (Plummer et al., 2015).

Referring expressions is the standard grounding objective, included in training of nearly all LVLMs. Given a natural language description (of a region), the model has to ground it to the correct image region. As is common practice, we also use the inverse task, that is, generation of the natural language description for the given image region.

Grounded captioning is the task of generating an image caption in which the locations of regions for mentioned objects are interleaved in the caption (see Figure 3 for examples). In theory, such explicit grounding is expected to result in closer adherence to the image content and reduce hallucinations.

Other grounding objectives have been proposed for LVLMs training, such as question answering with image regions in the input or output (Zhu et al., 2016); these, however, are outside the scope of our study, since we focus on the effects of grounding on hallucination primarily in free-form captioning.

Encoding regions. Different approaches exist for representing image regions for the LVLMs. Most commonly, regions are represented as bounding boxes using either (relative) coordinates in "plain text" (Liu et al., 2023b; Chen et al., 2023b; Bai et al., 2023; Wang et al., 2023b) (e.g., "$[0.10, 0.05, 0.64, 1.00]$"; the coordinates are treated as text and tokenized normally) or with learned embeddings corresponding to a fixed-size rasterization of the image (Peng et al., 2023; You et al., 2023; Pramanick et al., 2023). In this work, we adopt the former region representation, i.e., relative coordinates as text, as this does not introduce any additional trainable parameters to the model.

3 Measuring Object Hallucination

LVLM object hallucination is evaluated via two main protocols: (1) in QA-based evaluation, where models answer questions about object existence in the image (Li et al., 2023b) and (2) in open generation (usually image captioning) (Rohrbach et al.,

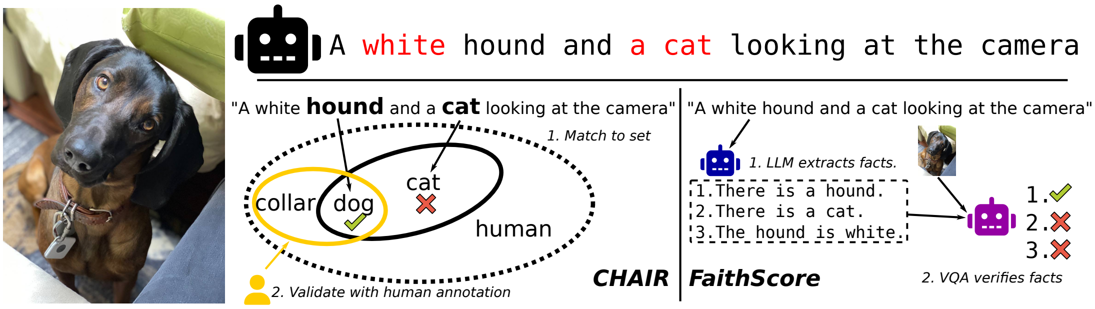


Figure 1: **CHAIR** and **FaithScore** are used to measure hallucinations in open caption generation with LVLMs. **CHAIR** relies on human object annotation (over a fixed set) to identify objects and check if they are hallucinated. **FaithScore** first uses an LLM to convert captions into facts which are then verified by a VQA model.

2018; Wang et al., 2023a; Jing et al., 2023). Measuring hallucination in the latter is arguably more indicative of models' tendency to hallucinate "in the wild", but it is also more difficult to devise automatic metrics. In contrast, QA-based evaluation is straightforward but is merely a proxy for actual hallucination in generative tasks.

QA-Based Hallucination Evaluation. POPE (Li et al., 2023b) is the de-facto standard benchmark for QA-based hallucination evaluation. Relying on images annotated for object detection (i.e., MSCOCO (Lin et al., 2014)), the benchmark consists of yes/no questions about object existence ("*Is there X in the image?*"). The negative questions – with objects *not* in the image – are generated in three different ways using: i) objects randomly selected from the total pool of objects that exist in the dataset (*random*); ii) the most frequently annotated objects in the dataset (*popular*); iii) objects with high co-occurrence to the image's actual objects (*adversarial*), as co-occurrence statistics are a common cause of hallucinations (Rohrbach et al., 2018; Biten et al., 2022; Li et al., 2023b; Zhou et al., 2023). The performance metric is accuracy, i.e., the percentage of correctly answered questions.

Open Hallucination Evaluation. We select CHAIR (Rohrbach et al., 2018) and FaithScore (Jing et al., 2023) (illustrated in Figure 1) to quantify hallucinations in open caption generation. The two metrics identify hallucinations in distinct manners. By adopting both, we mitigate the risk of our findings being merely an artifact of a single (imperfect) evaluation metric.

Both metrics also indirectly measure how *informative* and descriptive the generated captions are. As our results (§5) show, there exists a tradeoff between faithfulness/hallucinations and informativeness of the captions. We thus argue that the hallucination metrics should be contextualized with the measures of informativeness: factual yet non-informative captions are as useless as captions with a lot of hallucinated information.

CHAIR detects hallucinated objects using the set of 80 object classes from MSCOCO (Lin et al., 2014) with which the images are annotated. Words from the captions are matched—using string matching—against the class names (augmented with synonyms). The resulting list of matched objects is then cross-referenced against the gold list of annotated objects and all matched but not annotated objects are considered hallucinations. Two scores are produced over the dataset: (1) $CHAIR_i$ divides the total number of hallucinated objects across all captions with the total number of detected objects; (2) $CHAIR_s$ is the proportion of images in the dataset for which the caption contains at least one object hallucination. $CHAIR_s$ is less than ideal for longer captions as they are more likely to contain at least one hallucination: the binary caption-level measure could camouflage substantial differences in hallucination extent between models. Because of this, we adopt only $CHAIR_i$ in this work. Following Zhai et al. (2023a), we report the average number of matched objects per caption as well as the gold object coverage (i.e., the average percentage of annotated objects mentioned in the caption) as measures of informativeness.

CHAIR unfortunately comes with two major shortcomings. First, it is based on MSCOCO images and object annotations which are widely used in a range of derivative datasets leveraged for training LVLMs (Goyal et al., 2017; Kazemzadeh et al., 2014; Mao et al., 2016; Liu et al., 2023c). This makes LVLMs a priori less likely to hallucinate on MSCOCO images, which means that CHAIR is likely overly optimistic about (i.e., it underestimates) the amount of LVLM hallucination "in the wild". We thus propose to extend CHAIR to an out-of-distribution dataset, one that ideally also comes with a larger set of object classes. Second, CHAIR relies on exact string matching between caption words and synonyms of the object classes. Adapting vanilla CHAIR based on string matching to a larger set of object classes would, however, require significant manual effort: one has to (1) create a curated list of synonyms for all classes (without overlap between related classes) to correctly account for recall and (2) inspect examples and create special rules for edge cases to limit false positives (e.g., add 'baby X' synonyms to all animal classes in order not to falsely match the 'person' class). Addressing both issues simultaneously, we propose semantic matching between the caption and object classes as an alternative to string matching for large sets of object classes. Our extension, dubbed **CHAIR-MEN** (from **CHAIR** with **M**atching using **E**mbeddings of **N**oun phrases) (1) extracts all noun phrases from the generation,[1] (2) embeds the extracted phrases as well as classes names with a pretrained sentence encoder (Reimers and Gurevych, 2019)[2] and (3) makes matching decisions based on cosine similarity between obtained embeddings: to each noun phrase, we assign (i) the class amongst the image's objects with the most similar embedding, if cosine exceeds a threshold t_1, (ii) the class amongst the other objects (i.e., not present in the image) with the most similar embedding, if cosine exceeds a threshold t_2, or otherwise c) no object. Matching first only against the image's objects makes false negatives from a semantically related object not in the image less likely. We calibrate the thresholds ($t_1 = 0.73, t_2 = 0.78$) by trying to match the scores that CHAIR produces on MSCOCO, as an established measure for that dataset.

FaithScore (Jing et al., 2023), a model-based hallucination metric, is designed with finer-grained evaluation in mind: it does not only consider objects/entities but also other aspects that models can hallucinate about (specifically: color, relation, count, and 'other' attributes), without the need for human annotation. FaithScore computation is a 2-stage process that relies: (1) on an LLM to extracts 'atomic facts' from the generated text, phrasing them as statements ("There is a man") the factuality of which, in the context of the image, is then (2) verified with a VQA model ("Is the following statement correct?"). The final score is then simply the proportion of positive answers given by the VQA model. We additionally report the average number of facts produced by the LLM as a measure of informativeness of generated captions. The original work of Jing et al. (2023) relies on GPT-4 to extract facts but this is too expensive for our evaluation; instead, we use a smaller LLM[3] after verifying that it successfully follows task instructions. We use OFA (Wang et al., 2022) as the VQA model for FaithScore, as it is much faster and only marginally less accurate than Llava-1.5 (Liu et al., 2023b) according to Jing et al. (2023).

Caption Quality Metrics. We include the following metrics to monitor how grounding objectives affect the general caption quality: **CIDEr** (Vedantam et al., 2015) is a measure based on n-gram overlap with a set of reference captions. **CLIP-Score**, a reference-free metric, is the cosine similarity between the image and caption embeddings, produced by a CLIP model (Radford et al., 2021)[4].

4 Experimental Setup

We comprehensively analyze the effect of grounding objectives on LVLM hallucination. For the sake of transferability (and robustness) of our findings, the experimental core, namely the model architecture and training procedure, follows established practices as closely as possible. All model instances are trained according to the same protocol, that is, we control for everything other than the effect of grounding, i.e., inclusion/exclusion of grounding data in training. We primarily focus on measuring hallucination in open-ended image captioning as this, we argue, better reflects LVLM's hallucination in real-world applications; for completeness and comparison of evaluation protocols, we also perform the QA-based evaluation with POPE. We benchmark LVLMs for hallucinations in three different caption generation scenarios: (1) in standard image captioning, with expected caption length of 1-2 sentences (as in MSCOCO), (2) long (i.e., detailed, descriptive) caption generation, and (3) grounded image captioning (with standard length), where the LVLM is explicitly prompted to

[1] With spaCy v3 EN_CORE_WEB_SM

[2] BAAI/BGE-BASE-EN-V1.5 (Xiao et al., 2023)

[3] TEKNIUM/OPENHERMES-2.5-MISTRAL-7B which is based on Mistral-7B (Jiang et al., 2023); inference done with vLLM (Kwon et al., 2023) for speed

[4] We use VIT-B-16-SIGLIP-256 (Zhai et al., 2023b)

interleave region coordinates into the caption.

Evaluation Datasets. Despite the previously mentioned shortcomings, **MSCOCO** (Lin et al., 2014) remains the primary dataset for evaluating LVLM hallucination in the literature, both with QA-based and free-form generation metrics/protocols (Rohrbach et al., 2018; Li et al., 2023b). Hence, we include MSCOCO but complement it with the **Objects365** (**O365**) (Shao et al., 2019) dataset which comes with a much larger inventory of object classes (365 classes in total, including the 80 MSCOCO classes) and, consequently, more object annotations per image. We evaluate on 5000 and 5386 images from test portion of MSCOCO and validation portion of O365, respectively.[5]

For POPE, the QA-based hallucination metric, we generate two new test sets from O365, each with 1500 examples (matching the MSCOCO POPE examples): `O365/COCO` uses only the 80 classes from MSCOCO, and `O365/non-COCO` relies on the remaining 285 classes.

LVLM Architecture. We adopt the architecture typical for most LVLMs: (1) images are encoded by an image encoder, (2) projected by an alignment module into the LLM embedding space, and (3) prepended to the embeddings of textual tokens (Li et al., 2023a). The alignment module in our experiments is a resampler (Li et al., 2023a; Bai et al., 2023; Alayrac et al., 2022), a type of Transformer (Vaswani et al., 2017) that learns to encode the visual information from the image in a set of trainable query embeddings; specifically, we use a 3-layer perceiver-resampler (Alayrac et al., 2022). The number of query tokens (32 in our experiments) is a lot smaller than the number of visual embeddings at the output of the image encoder (256), which leads to more efficient training.[6] We use a frozen SigLIP (Zhai et al., 2023b) (VIT-SO400M-14) as the image encoder and Vicuna 1.5 7B (Chiang et al., 2023) as LLM. The original LLM parameters are frozen and 4-bit quantized (Dettmers et al., 2023); instead of direct LLM updates, we learn the LoRA adapters (Hu et al., 2022) for all Transformer matrices (with $r = 32, \alpha = 64$).

Training Mix. LVLMs are generally trained on a mix of tasks and datasets. The mix we adopt reflects the main goal of our empirical study: investigating how training with grounding affects LVLMs regarding hallucination in free-form caption generation and comparing it to hallucinations in QA. Our mix thus includes the following tasks:

1. *Standard image captioning*: we train on MSCOCO and 1M examples sampled from CC3 (Sharma et al., 2018) and SBU (Ordonez et al., 2011) (with synthetic captions produced by Li et al. (2022)) for a total of 1.4M image-caption pairs.
2. *Long captioning*: We use LLAVA-DETAILED (Liu et al., 2023c) with 23k long captions generated by GPT-4 on the basis of (short) MSCOCO reference captions and gold object annotations.
3. *VQA*: We select from VQAv2 (Goyal et al., 2017) all 170k yes/no questions. VQA is only added to the training mix for the QA-based hallucination evaluation protocol (i.e., POPE). [7]
4. *Referring expressions* (see §2): we combine RefCOCO (Kazemzadeh et al., 2014; Mao et al., 2016) (320k examples) and Visual Genome (Krishna et al., 2017) (we sample 1M examples).
5. *Grounded captions* (see §2): we use Flickr30k-Entities (Plummer et al., 2015) (150k examples).

We name our LVLM model variants based on their respective training mix. The `Base` LVLM has been trained only on non-grounding tasks (1-3); addition of the referring expressions and grounded captioning tasks is indicated with `+RE` and `+GC`, respectively. For brevity, we provide further training and inference details in the Appendix A.

5 Results

We now report the observed hallucination effects under both protocols: in free-form captioning and in QA-based hallucination evaluation (as indicated by POPE). All results are averages over three runs with different random seeds. The reported CHAIR results correspond to our CHAIR-MEN variant; we report the results obtained with the vanilla CHAIR based on string matching in Appendix B.

QA Hallucinations with POPE. Table 1 summarizes the hallucination results in a QA-based evaluation protocol with POPE. Generally, grounding, based both on referring expressions (`+RE`)

[5] We have additionally considered Open Images (Kuznetsova et al., 2020), Visual Genome (VG) (Krishna et al., 2017), and LVIS (Gupta et al., 2019) as datasets with object annotations but ultimately decided against their inclusion due to insufficient object coverage in annotations (i.e., not all objects are annotated in every image).

[6] Using visual embeddings directly, without the resampler projection, like in Llava (Liu et al., 2023c) would have more than doubled our training time.

[7] Without VQA in the training mix, the LVLM does not follow the POPE task instruction.

Model	MSCOCO rand.	MSCOCO pop.	MSCOCO adv.	O365/COCO rand.	O365/COCO pop.	O365/COCO adv.	O365/non-COCO rand.	O365/non-COCO pop.	O365/non-COCO adv.
Base	86.37	81.82	78.18	79.31	71.44	66.72	76.51	69.09	61.37
+RE	86.72	83.98	80.28	**81.40**	**73.92**	**68.82**	**77.88**	**72.33**	**64.44**
+GC	86.38	83.89	79.88	79.11	71.72	67.11	76.98	69.49	62.18
+RE+GC	**87.03**	**84.43**	**80.98**	80.49	73.50	68.37	76.11	70.66	63.32

Table 1: POPE results (accuracy) for MSCOCO, O365/COCO (using the 80 MSCOCO object classes), and O365/non-COCO (remaining 285 classes) for random, popular, and adversarial example sets.

	Model	CIDEr↑	CLIPS.↑	#Words	$CHAIR_i$ ↓	Coverage↑	Objects	FaithScore↑	Facts
MSCOCO	Base	62.46	13.04	16.35	5.28	60.50	1.98	80.55	7.34
	+RE	6.51	**13.37**	**30.03**	7.83	**66.59**	**2.88**	78.78	**10.68**
	+GC	**78.05**	12.84	15.04	**4.52**	60.01	1.90	**81.88**	7.01
	+RE+GC	53.49	13.09	20.00	6.43	62.53	2.27	80.59	8.26
Objects365	Base	—	12.46	15.64	16.89	33.54	2.62	77.07	7.56
	+RE	—	**12.74**	**28.70**	21.19	**39.24**	**3.94**	76.50	**11.34**
	+GC	—	12.33	14.68	**15.27**	32.98	2.47	**78.72**	7.23
	+RE+GC	—	12.45	19.34	18.28	35.82	3.03	77.91	8.61

Table 2: Results on standard image captioning. CIDEr and CLIPScore indicate general caption quality; $\texttt{CHAIR}_i$ and FaithScore reflect hallucination, whereas (average number of) #Words, CHAIR Coverage and Objects, and (number of FaithScore) Facts aim to quantify informativeness. **Bold**: the best value in column; underline: the worst.

and grounded captions (+GC) seems to lead to performance gains, i.e., hallucination reduction (1-3 points on popular and adversarial subsets). RE brings more substantial gains than GC, not only on MSCOCO data (due to training on RefCOCO) but also on out-of-distribution images, i.e., on O365/non-COCO; combining the two grounding objectives, however, brings further gains only on MSCOCO. These results generally align well with the findings from prior work on grounding-based LVLM training (Chen et al., 2023b; You et al., 2023). Grounding objectives thus seem to improve the fine-grained image understanding, at least with respect to object existence. We next investigate whether these gains translate to hallucination reduction in free-form caption generation.

Standard Captions. The performance of the LVLM variants on standard image captioning is shown in Table 2. Referring expressions (+RE), compared to the Base model, doubles the average caption length from 16 to 30 words. The additional content seems informative according to CHAIR Objects count, CHAIR Coverage, and FaithScore Facts. Unfortunately, the longer captions exhibit not just an absolute increase in hallucinated content but also a *relative* increase, since both $CHAIR_i$ and FaithScore are length-normalized metrics.

The effect is different for training on grounded captions (but prompting at inference for standard captions without bounding boxes): +GC leads to slightly better $CHAIR_i$ and FaithScore, but it also slightly reduces the informativeness of the captions. Interestingly, GC seems to 'counteract' the effect of RE as their combination (+RE+GC) leads to substantially shorter captions than +RE alone.

As for the common captioning metrics, we observe that CIDEr prefers shorter captions, whereas CLIPScore slightly prefers the longer, more descriptive captions. Finally, we consider a fine-grained analysis of FaithScore in Appendix C.

Grounded Captions. Intuitively, we would have expected that training to generate grounded captions, would prompt the model to only generate objects that it can actually ground in the image. Looking at the results in Table 3, we see that, while CHAIR-based metrics indeed suggest a lower level of hallucination (in comparison to Table 2), the FaithScore accuracy does not improve (in fact, it even slightly worsens; compare against the corresponding values for standard captioning from Table 2). We also note generating grounded captions leads to a general reduction in informativeness, e.g., lower averages of CHAIR Objects and FaithScore Facts (compare, again, against corresponding values in Table 2). Similarly, combining the two grounding objectives in training (+RE+GC) leads to slightly more hallucinative grounded captions according to FaithScore.

These results add to the conclusion that grounding objectives generally fail to reduce hallucination

	Model	CIDEr↑	CLIPS.↑	#Words	$CHAIR_i$ ↓	Coverage↑	Objects	FaithScore↑	Facts
MSCOCO	`+GC`	79.44	12.65	14.89	3.23	52.82	1.57	78.93	6.60
	`+RE+GC`	88.04	12.47	13.54	3.15	51.58	1.51	80.51	6.14
Objects365	`+GC`	—	12.10	14.62	12.78	28.63	2.00	77.03	6.77
	`+RE+GC`	—	11.94	13.45	12.27	27.84	1.89	77.76	6.33

Table 3: Performance on grounded image captioning. CIDEr and CLIPScore indicate overall caption quality; `CHAIR`$_i$ and `FaithScore` reflect hallucination, whereas (average number of) #Words, CHAIR Coverage and Objects, and (number of FaithScore) Facts aim to quantify informativeness.

Model	#Words	$CHAIR_i$ ↓	Coverage↑	Objects
`Base`	103.49	22.79	77.78	6.53
`+RE`	103.15	22.75	78.51	6.50
`+GC`	106.17	23.13	78.25	6.62
`+RE+GC`	104.58	22.65	78.23	6.54

Table 4: Results for long captions on MSCOCO. We report the average number of words and CHAIR metrics. Results with FaithScore and on O365 are qualitatively the same so we omit them for brevity.

in caption generation. A qualitative look (see §6) reveals that models trained with grounding objectives still incorrectly describe objects or fabricate them entirely (with bounding boxes). We also observe that on O365, more than half of the hallucinated objects (according to CHAIR) in the grounded captions are also hallucinated in respective standard captions; this suggests that causes beyond insufficient grounding are behind hallucination.

Long Captions. Table 4 shows long captioning results. For brevity, we only report the results for MSCOCO with CHAIR(-MEN): for O365 and FaithScore the results are qualitatively the same. Overall, the differences between model variants are negligible. We believe that, due to the small number of examples in LLAVA-DETAILED (only 23k, much less than for other training tasks) and their formulaic style (generated by GPT-4), all LVLM variants overfit to this style. A brief inspection of distributions of caption length supports this: all models nearly perfectly follow the training distribution. The grounding objectives (`+RE` and `+GC`) thus does not seem to affect long captions, in contrast to standard captions. This again questions the extent to which improved fine-grained image understanding from grounding actually transfers to hallucination reduction in open generation.

6 Qualitative Analysis

Standard Captions. Figure 2 shows captions generated by our different models. As already indicated by the automatic metrics, `Base` and `+GC` generate shorter captions. `+RE`-trained models produce longer and more detailed captions but they are also more likely to fabricate details.

Grounded Captions. We show examples for grounded captioning in Figure 3. Grounded captions are generally shorter , which coincides with a decrease in hallucinations but also a decrease in informativeness. However, the grounding itself does not seem to prevent the model from hallucinating: in one example, one correctly grounded kayak falsely becomes *yellow*; in the second example, the caption mentions *three zebras* yet only two are grounded; similarly, the correct bounding box is generated for the *wildebeest*, but it is incorrectly called a *gazelle*. The standard caption also falsely describes the *wildebeest* as *antelopes*, pointing to a cause other than insufficient grounding. In the third example, *spices*, together with the bounding box, are fully hallucinated by the grounded model.

7 Related Work

Large Vision-Language Models. LVLMs are essentially Large Language Models (LLMs) (Brown et al., 2020; Touvron et al., 2023; OpenAI, 2023; Jiang et al., 2023) extended to "understand" visual input. Recent models have shown an impressive understanding of images (OpenAI, 2023; Anil et al., 2023; Li et al., 2023a; Dai et al., 2023a; Liu et al., 2023c; Bai et al., 2023; Fini et al., 2023; Zhu et al., 2023; Laurençon et al., 2023; Geigle et al., 2023; Wang et al., 2023b) and a range of models have been proposed specifically for grounding and referring (Chen et al., 2023b; You et al., 2023; Pramanick et al., 2023; Zhang et al., 2023a; Peng et al., 2023; Chen et al., 2023a; Zhao et al., 2023a).

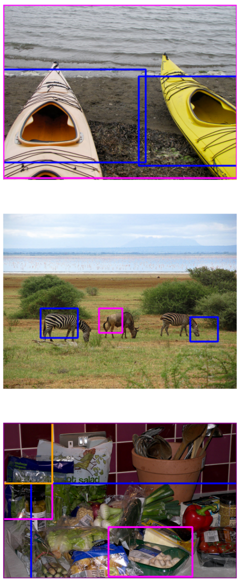
Measuring Object Hallucinations. A range of hallucination metrics have been proposed: CHAIR (Rohrbach et al., 2018) identifies hallucinated objects by checking captions (via string matching) against a set of annotated objects (i.e., MSCOCO). Wang et al. (2023a) fine-tune an LLM to identify

Base: Three toy horses are standing in a row, one is brown and the other two are white with black.
+GC: Three horses standing next to each other in front of a wall.
+RE: A white, a brown and a tan horse are standing in front of a wall with polka dots on it while one is wearing a bridle and the other two have saddles on their backs.
+RE+GC: Three toy horses are standing in a row. One is brown, one is white and the third has a black mane and tail with a white face and legs and body.

Base: A dog is laying on the ground next to a pool with his head in the water and paws up against the side of the pool.
+GC: A dog is laying on the ground next to a swimming pool and drinking from the water.
+RE: A dog with a long, golden coat is laying on the ground in front of a pool and has its head down by the edge of the water as it looks at something below.
+RE+GC: A dog is laying on the ground next to a pool and looking at something in its mouth.

Base: A black dog with a plate of broccoli in front of it and another bowl on the floor.
+GC: A black dog standing next to a plate of broccoli.
+RE: A black dog with a white face and brown ears, standing on a hardwood floor in front of a table that has a plate of broccoli on it and a bowl next to the plate, both containing food.
+RE+GC: A black dog with a plate of broccoli in front of it and another plate on the floor behind it.

Figure 2: Qualitative examples for standard captions generated by the different models. Hallucinations in red.



Standard: Two kayaks are sitting on the beach, one yellow and one orange.
Grounded: Two yellow kayaks are sitting on the beach .

Standard: A herd of zebra and a group of antelope grazing in the grass near water with mountains on horizon, with blue sky above and clouds in the distance.
Grounded: Three zebras and a gazelle graze in the grass near a body of water.

Standard: A kitchen counter with a pot, pans and various vegetables.
Grounded: A kitchen counter with a variety of vegetables, spices and cooking supplies.

Figure 3: Qualitative examples of +RE+GC for standard and grounded captioning. Hallucinations are underlined in red. Coordinates after the colored phrases are removed for readability and shown on the image.

hallucinatory captions through comparison with reference captions; FaithScore (Jing et al., 2023), a reference-free approach, uses an LLM to extract verifiable facts and then tests these facts with a VQA model. POPE (Li et al., 2023b) indirectly measures hallucination with questions about object existence: while a good test of image understanding , which may indicate the extent of models' tendency to hallucinate, it is not a direct measure of hallucination in open-ended captioning.

Hallucination Mitigation. A range of approaches have been proposed to mitigate hallucination: Biten et al. (2022); Dai et al. (2023b); Zhai et al. (2023a) propose adaptions to the training data and objectives. Liu et al. (2023a); Gunjal et al. (2023); Zhao et al. (2023b); Yu et al. (2023) use reinforcement-learning methods to reduce hallucinations in model output. Leng et al. (2023); Huang et al. (2023) propose (training-free) decoding methods that mitigate hallucinations. Zhou et al. (2023); Yin et al. (2023) create pipeline approaches that post-hoc clean the generated text from hallucinated content. Finally, for QA hallucinations, researchers have created robust instruction data (Liu et al., 2023a), VQA examples (Hu et al., 2023), and additional benchmarks (Lu et al., 2023).

8 Conclusion

Object hallucinations remain one of the main obstacles to wide-range adoption of LVLMs. Prior work suggested that grounding objectives like referring expressions reduce hallucinations but the empirical support for this claim is confined to QA-based evaluation. In this work, we performed an in-depth analysis of the effects of grounding objectives in LVLM training on hallucination in open image captioning. While our extensive experiments confirm that grounding objectives improve fine-grained image understanding and show that they lead to substantially more detailed and informative captions, we find little evidence that they actually reduce hallucination in open caption generation; on the contrary, they often even increase the amount of hallucinated content. Our findings warrant efforts towards hallucination mitigation in image captioning that go well beyond object grounding alone.

9 Limitations

There are two main limitations to our analysis. First, while we aim for a comprehensive analysis of the effects of different training objectives and task mixes on downstream hallucination (for example, we execute multiple runs for each model variant and average the results), there are a number of modeling decisions that we had to fix (i.e., we could not explore other variants)—primarily w.r.t. to the architecture of the LVLM— due to a limited computational budget. One could, inter alia, consider a different image encoder, a different/larger LLM, and/or alignment modules other than perceiver-resampler. Additionally, due to our limited computational budget, we train our models on less data and for fewer steps than a lot of other work that trains LVLMs (e.g. Chen et al. (2023b); Liu et al. (2023b); Bai et al. (2023)); we thus cannot rule out that a reduction in hallucination due to grounding objectives might *emerge* at some larger scale of grounding training.

Second, our findings are (modulo anecdotal evidence from manual qualitative analysis of a limited number of examples) based on reliance on imperfect automatic metrics. While this is a common practice in related work as well, we increase the likelihood of the robustness of our findings and conclusions by employing two mutually complementing hallucination quantification metrics, CHAIR and FaithScore (see §3), as well as additionally proposing a semantic extension to CHAIR (CHAIR-MEN, see §3).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. *CoRR*, abs/2204.14198. ArXiv: 2204.14198.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *CoRR*, abs/2312.11805. ArXiv: 2312.11805.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *CoRR*, abs/2308.12966. ArXiv: 2308.12966.

Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. 2022. Let there be a clock on the beach: Reducing Object Hallucination in Image Captioning. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2473–2482. IEEE.

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. *arXiv:2005.14165 [cs]*. ArXiv: 2005.14165.

Emanuele Bugliarello, Laurent Sartran, Aishwarya Agrawal, Lisa Anne Hendricks, and Aida Nematzadeh. 2023. Measuring Progress in Fine-grained Vision-and-Language Understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 1559–1582. Association for Computational Linguistics.

Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. 2023a. Position-Enhanced Visual Instruction Tuning for Multimodal Large Language Models. *CoRR*, abs/2308.13437. ArXiv: 2308.13437.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023b. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *CoRR*, abs/2306.15195. ArXiv: 2306.15195.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023a. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *CoRR*, abs/2305.06500. ArXiv: 2305.06500.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. 2023b. 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, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2128–2140. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. *CoRR*, abs/2305.14314.

Enrico Fini, Pietro Astolfi, Adriana Romero-Soriano, Jakob Verbeek, and Michal Drozdzal. 2023. Improved baselines for vision-language pre-training. *CoRR*, abs/2305.08675. ArXiv: 2305.08675.

Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavas. 2023. mBLIP: Efficient Bootstrapping of Multilingual Vision-LLMs. *CoRR*, abs/2307.06930. ArXiv: 2307.06930.

Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6325–6334.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2023. Detecting and Preventing Hallucinations in Large Vision Language Models. *CoRR*, abs/2308.06394. ArXiv: 2308.06394.

Agrim Gupta, Piotr Dollár, and Ross B. Girshick. 2019. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5356–5364. Computer Vision Foundation / IEEE.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023. CIEM: Contrastive Instruction Evaluation Method for Better Instruction Tuning. *CoRR*, abs/2309.02301. ArXiv: 2309.02301.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. *CoRR*, abs/2311.17911. ArXiv: 2311.17911.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *CoRR*, abs/2310.06825. ArXiv: 2310.06825.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. FAITHSCORE: Evaluating Hallucinations in Large Vision-Language Models. *CoRR*, abs/2311.01477. ArXiv: 2311.01477.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 787–798. ACL.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. *CoRR*, abs/1909.05858. ArXiv: 1909.05858.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vision*, 123(1):32–73. Place: USA Publisher: Kluwer Academic Publishers.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The Open Images Dataset V4. *Int. J. Comput. Vis.*, 128(7):1956–1981.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles, SOSP 2023, Koblenz, Germany, October 23-26, 2023*, pages 611–626. ACM.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. 2023. OBELISC: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. *CoRR*, abs/2306.16527. ArXiv: 2306.16527.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. *CoRR*, abs/2311.16922. ArXiv: 2311.16922.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *CoRR*, abs/2301.12597. ArXiv: 2301.12597.

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. 2023b. Evaluating Object Hallucination in Large Vision-Language Models. *CoRR*, abs/2305.10355. ArXiv: 2305.10355.

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.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Aligning Large Multi-Modal Model with Robust Instruction Tuning. *CoRR*, abs/2306.14565. ArXiv: 2306.14565.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved Baselines with Visual Instruction Tuning. *CoRR*, abs/2310.03744. ArXiv: 2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual Instruction Tuning. *CoRR*, abs/2304.08485. ArXiv: 2304.08485.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl J. Yang, and Jie Yang. 2023. Evaluation and Mitigation of Agnosia in Multimodal Large Language Models. *CoRR*, abs/2309.04041. ArXiv: 2309.04041.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016. Generation and Comprehension of Unambiguous Object Descriptions. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 11–20. IEEE Computer Society.

OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774. ArXiv: 2303.08774.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain*, pages 1143–1151.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding Multimodal Large Language Models to the World. *CoRR*, abs/2306.14824. ArXiv: 2306.14824.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649.

Shraman Pramanick, Guangxing Han, Rui Hou, Sayan Nag, Ser-Nam Lim, Nicolas Ballas, Qifan Wang, Rama Chellappa, and Amjad Almahairi. 2023. Jack of All Tasks, Master of Many: Designing General-purpose Coarse-to-Fine Vision-Language Model. _eprint: 2312.12423.

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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. 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, Brussels, Belgium, October 31 - November 4, 2018*, pages 4035–4045. Association for Computational Linguistics.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. 2019. Objects365: A Large-Scale, High-Quality Dataset for Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8429–8438. IEEE.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning Large Multimodal Models with Factually Augmented RLHF. *CoRR*, abs/2309.14525. ArXiv: 2309.14525.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971. ArXiv: 2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. 2023a. Evaluation and Analysis of Hallucination in Large Vision-Language Models. *CoRR*, abs/2308.15126. ArXiv: 2308.15126.

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.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023b. CogVLM: Visual Expert for Pretrained Language Models. *CoRR*, abs/2311.03079. ArXiv: 2311.03079.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighof. 2023. C-Pack: Packaged Resources To Advance General Chinese Embedding. *CoRR*, abs/2309.07597. ArXiv: 2309.07597.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination Correction for Multimodal Large Language Models. *CoRR*, abs/2310.16045. ArXiv: 2310.16045.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and Ground Anything Anywhere at Any Granularity. *CoRR*, abs/2310.07704. ArXiv: 2310.07704.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2023. RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback. *CoRR*, abs/2312.00849. ArXiv: 2312.00849.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2023a. HallE-Switch: Rethinking and Controlling Object Existence Hallucinations in Large Vision Language Models for Detailed Caption. *CoRR*, abs/2310.01779. ArXiv: 2310.01779.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023b. Sigmoid Loss for Language Image Pre-Training. *CoRR*, abs/2303.15343. ArXiv: 2303.15343.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. 2023a. GPT4RoI: Instruction Tuning Large Language Model on Region-of-Interest. *CoRR*, abs/2307.03601. ArXiv: 2307.03601.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *CoRR*, abs/2309.01219. ArXiv: 2309.01219.

Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2023a. Bubogpt: Enabling visual grounding in multi-modal llms. *CoRR*, abs/2307.08581.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023b. Beyond Hallucinations: Enhancing LVLMs through

Hallucination-Aware Direct Preference Optimization. *CoRR*, abs/2311.16839. ArXiv: 2311.16839.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. *CoRR*, abs/2310.00754. ArXiv: 2310.00754.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *CoRR*, abs/2304.10592. ArXiv: 2304.10592.

Yuke Zhu, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded Question Answering in Images. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4995–5004. IEEE Computer Society.

Task	Prompt
Standard Caption	Briefly describe the image.
Long Caption	Describe the image in detail.
Grounded Caption	Describe the image and include the bounding box coordinates for every mentioned object.
VQA (POPE)	QUESTION Answer with yes or no.
Referring Expression	Give the bounding box coordinates for the region described as "DESCRIPTION".
Referring Generation	Briefly describe the region [x1, y1, x2, y2].

Table 5: Prompts used for training and inference.

A Training and Details

All models were trained on a single NVIDIA RTX3090s card, with training duration ranging between 4 and 7 GPU days, depending on the training task mix. We train for one epoch (on the concatenation of corpora from all tasks, as all tasks are—from the low-level technical point of view—instances of causal language modeling, i.e., next token prediction) with AdamW optimizer (Loshchilov and Hutter, 2019), learning rate 2e-4, weight decay 0.01, batch size 128 (achieved with gradient checkpointing and accumulation), and a cosine schedule.

For generation (i.e., inference), we use greedy decoding with a repetition penalty (Keskar et al., 2019) of 1.15 to avoid degenerative repetitions in long caption generation. We use one fixed prompt per task (see Table 5) both in training and at inference (for the subset of tasks on which we evaluate).

We encode bounding boxes with 2 significant digits (, e.g., $[0.10, 0.05, 0.64, 1.00]$). For grounded captions where multiple bounding boxes are needed (e.g., for something like "three zebras"), we follow Plummer et al. (2015) and combine the coordinates with semicolons in the same brackets (, e.g., $[0.10, 0.05, 0.64, 1.00; 0.50, 0.15, 0.64, 1.00]$).

B CHAIR and CHAIR-MEN

We report results based on our CHAIR-MEN approach in the main paper. In the following, we compare them against vanilla CHAIR results based on the string matching method. In Table 6, we report string-matching CHAIR results for MSCOCO, which can be compared to Table 2 (standard captions), Table 3 (grounded captions), and Table 4 (long captions).

Model	Coverage↑	$CHAIR_i\downarrow$	Objects
`Base`	63.85	6.77	2.04
`+RE`	67.90	10.96	2.87
`+GC`	63.43	5.97	1.96
`+RE+GC`	65.42	8.40	2.31

(a) MSCOCO Standard Captions

Model	Coverage↑	$CHAIR_i\downarrow$	Objects
`+GC`	57.52	4.23	1.67
`+RE+GC`	56.44	3.96	1.60

(b) MSCOCO Grounded Captions

Model	Coverage↑	$CHAIR_i\downarrow$	Objects
`Base`	78.97	25.07	6.88
`+RE`	80.26	25.00	6.85
`+GC`	79.96	25.77	7.05
`+RE+GC`	79.90	25.39	6.92

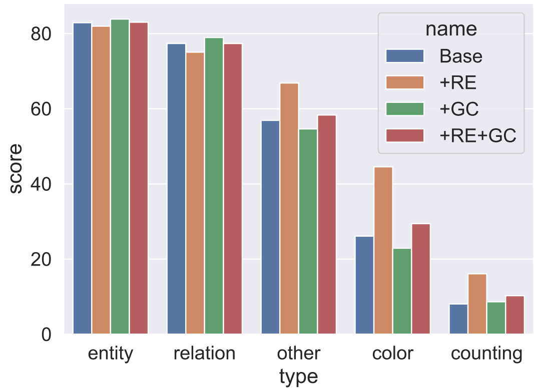
(c) MSCOCO Long Captions

Table 6: CHAIR results for MSCOCO using the classic string-matching approach.

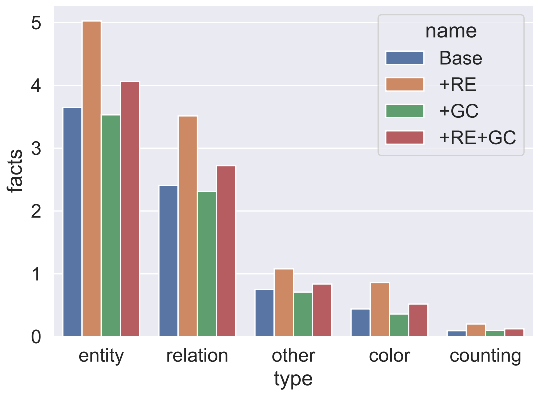
We find that results with CHAIR-MEN are highly proportional to CHAIR: while $CHAIR_i$ and the number of overall objects found (along with the coverage) are slightly lower with CHAIR-MEN, the ranking between models are the same. This validates CHAIR-MEN as an alternative approach for identifying hallucinated objects and opens up the extension to other datasets like Objects365.

C Additional Results

Fine-grained Faithscore. Figure 4 offers a fine-grained analysis of different hallucination types, as predicted by FaithScore. While *entity* and *relation* hallucinations rates are similar across models, training with referring expressions (`+RE`) appears to greatly reduce hallucination w.r.t. *counting*, *color* (and, to a lesser extent, *other* attributes): the `+RE` model nearly doubles the the accuracy of the `Base` model. This is noteworthy because the number of *color/counting* facts also nearly doubles for `+RE`; this counters the a priori likelihood of having more hallucinations on more generated facts.



(a) Scores (accuracy) by type.



(b) Average number of facts by type.

Figure 4: Fine-grained look at the different types of hallucinations of FaithScore (standard MSCOCO captions). Results on Objects365 are qualitatively the same.