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). LVLMs, however, often *hallucinate* and produce captions that mention concepts that cannot be found in the image. These hallucinations erode the trustworthiness of LVLMs and are arguably among the main obstacles to their ubiquitous adoption. Recent work suggests that addition of grounding objectives—those that explicitly align image regions or objects to text spans—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 over three backbone LLMs reveal that grounding objectives have little to no effect on object hallucination in open caption generation.

1 Introduction

Large Vision-Language Models (LVLMs) have recently displayed impressive image understanding abilities (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 *object hallucination* in which the LVLMs—similar to "general" hallucination of LLMs (Zhang et al., 2023b)—"invent" objects (or attributes of or relations between objects) not present in the image.

A range of methods have recently been proposed to address LVLM hallucination such as modified decoding strategies (Leng et al., 2023; Huang et al., 2023), post-hoc removal of hallucinated content (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 of these approaches, however, either increase inference cost or need expensive additional training and/or data, impeding 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: region-level objectives demand finer-grained image understanding than the 'global' image captioning (*de facto* the main training objective of LVLMs), as demonstrated in visiolinguistic compositionality (Bugliarello et al., 2023). Such objectives should thus, intuitively, discourage models from generating content they cannot ground in the image. Intuition aside, the empirical support for the claim that grounding objectives reduce LVLM hallucination is weak and mainly limited to question-answering (QA) style of evaluation in which the model is explicitly asked about existence of objects in an image (Li et al., 2023b); we argue that this evaluation protocol poorly aligns with real-world *free-form* text generation tasks—primarily open image captioning—for which there is no empirical evidence yet that object grounding reduces hallucination.

Contributions. In this work, we perform the first comprehensive analysis of the effects that grounding objectives have on LVLM object hallucination in open (i.e., free-form) image captioning, addressing the shortcomings of existing hallucination evaluation protocols. Concretely, we measure the effect of adding two popular grounding objectives as additional objectives to standard image captioning-based training of LVLMs: (1) the *referring expressions (RE)* objective asks the model to generate the bounding box of the region that corresponds to a textual description and vice versa; whereas (2) the *grounded captioning (GC)* objective demands that the model generates image descriptions with interleaved (relative coordinates of) bounding boxes for mentioned objects. We then compare the extent of hallucination for LVLM variants trained with and without these grounding objectives. To this end, we compare the hallucination measures based on question answering (QA) (Li et al., 2023b) against free-form metrics for open captioning (Rohrbach et al., 2018; Jing et al., 2023). Critically, observing that (1) existing evaluation measures and protocols (Rohrbach et al., 2018; Li et al., 2023b) rely on MSCOCO (Lin et al., 2014) and (2) MSCOCO data is part of the training mix for most LVLMs, we argue that existing measures are likely to underestimate LVLM hallucinate; we thus extend our hallucination evaluation protocol to out-of-distribution data that LVLMs will not have seen in training.

Findings. Our experiments with three different LLM backbones show that, under a sound evaluation protocol, including grounding objectives—referring expressions and grounded captioning—to LVLM training has little to no effect on object hallucination, both in QA-based evaluation and open-ended captioning. Enforcing generation of *grounded captions* at inference time, on the other hand, slightly reduces object hallucinations but the effect is small and comes at the cost of (slight) reduction in caption detailedness. A qualitative inspection of grounded captions also confirms that forcing model to generate a bounding box for mentioned objects most often does not prevent it from hallucinating content. In sum, we find that grounding objectives fail to meaningfully reduce LVLM hallucination, calling for novel methodological proposals towards hallucination reduction.

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, in the form of a bounding box and predict corresponding language expressions or produce such regions as output. Many recent LVLMs have been trained with grounding tasks in their training mix alongside standard tasks like captioning and VQA (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 part of LVLM training: referring expressions (Kazemzadeh et al., 2014) and grounded captioning (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 2 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, because 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 with the tokenizer of the corresponding LLM) or with learned embeddings that correspond 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 avoids introducing 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 gener-

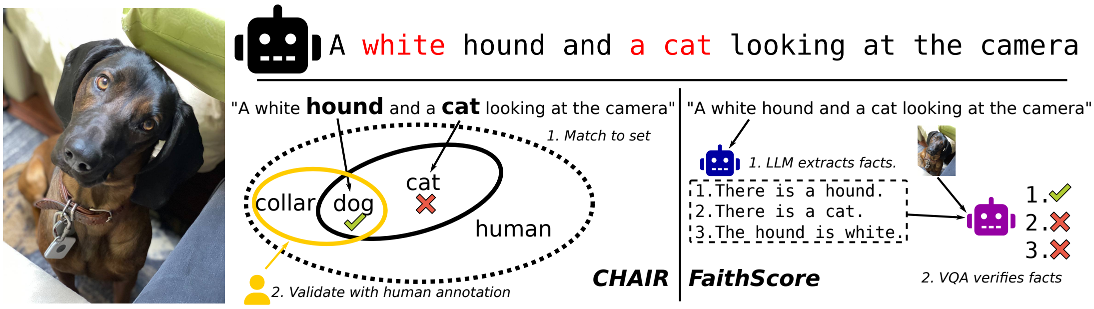

Figure 1: **CHAIR** and **FaithScore** are used to measure hallucinations in open caption generation with LVLMs. **CHAIR** relies on human object annotation (over a fixed set) to identify objects and check if they are hallucinated. **FaithScore** first uses an LLM to convert captions into facts which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

ation, usually image captioning (Rohrbach et al., 2018; Wang et al., 2023a; Jing et al., 2023). The latter is arguably more indicative of models' tendency to hallucinate "in the wild" (i.e., in various real-world applications) but it is also a more difficult setup for automatic evaluation. In contrast, QA-based evaluation is straightforward, but an untested 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 with objects from MSCOCO (Lin et al., 2014), the benchmark consists of *yes/no* questions about object existence ("*Is there X in the image?*"). The negative questions—about 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 focus on two popular meatrics for quantifying hallucination in open caption generation: CHAIR (Rohrbach et al., 2018) and FaithScore (Jing et al., 2023), illustrated in Figure 1). The two metrics identify hallucination in different ways: by complementing them with one another, we mitigate the risk of our findings merely being an artifact of a single (imperfect) evaluation metric. Both metrics can also indirectly quantify how *informative* and descriptive the generated captions are. As our result will show (§5), there exists a tradeoff between faithfulness/hallucination and informativeness of the captions. We thus argue that the hallucination metrics should be contextualized with the measures of informativeness: factually correct but uninformative captions are as undesired as captions with 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 exact 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; such a binary caption-level measure would hide potentially substantial differences in hallucination rates between models. Because of this, we adopt only CHAIR_i in this work. Following Zhai et al. (2023a), we additionally 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 caption *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 synonym sets of the object classes. Adapting vanilla CHAIR based on string matching to a larger set of object classes would, however, require significant manual effort, as one would have to (1) create a curated list of synonyms for all new 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 'X' 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 (iii) no object class. 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 vanilla 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 extract 'atomic facts' from the generated text, phrasing them as statements (e.g., *"There is a man"*) the factuality of which, in the context of the image, is then (2) verified with a VQA model (question: "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. Next to the hallucination measures, we add the following two standard 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., 2021a).[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, our 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 during 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 two different caption generation scenarios: (1) in *standard* image captioning, with expected caption length of 1-2 sentences (as in MSCOCO), and (2) *grounded* image captioning (with standard length), where the LVLM is explicitly prompted to interleave region

[1] With spaCy v3 EN_CORE_WEB_SM

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

[3] Llama3-8B-Instruct (AI@Meta, 2024); inference done with vLLM (Kwon et al., 2023) for speed

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

coordinates into the caption. In the Appendix B, we also provide results for *long* (i.e., detailed, descriptive) caption generation.

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). We thus 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 the POPE evaluation, we generate two new test sets from O365, each with 1500 examples (matching MSCOCO POPE): `O365/COCO` uses only the 80 classes from MSCOCO, and `O365/non-COCO` utilizes the remaining 285 classes.

LVLM Architecture. We adopt the typical LVLM architecture: (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 (Liu et al., 2023b). For the alignment module, we adopt as default the projection by Chu et al. (2024), which uses a 2-layer MLP followed by a pooling layer. We also experiment with a resampler (Li et al., 2023a; Bai et al., 2023; Alayrac et al., 2022), which 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) with 32 query tokens. We leverage the OpenAI CLIP ViT-L/14-224 (Radford et al., 2021b) as the image encoder. We experiment with three different LLM backbones: Vicuna 1.5 7B (Chiang et al., 2023), Llama-3 8B (instruct) (AI@Meta, 2024), and Phi-3-mini (Abdin et al., 2024). The 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 parameter matrices of the LLM.

[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 gold 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).

Pre-Training. We pre-train the alignment module—and only the alignment module (all other parameter frozen)—on image-caption data. For this, we use the 560k examples from Liu et al. (2023b).

Training Mix. LVLMs are generally instruction-trained on a mix of tasks and datasets. The mix we adopt reflects the main goal of our study: to isolate the effect of grounding objectives on LVLMs hallucination. We thus include the following tasks:
1. Standard image captioning: we train on the MSCOCO captions (400k examples);
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 sake of QA-based hallucination evaluation with POPE;[6]
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 320k examples);
5. Grounded captioning (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. By default, we use the pooled MLP projection from Chu et al. (2024) for all models. Additionally, we train a Vicuna-based model with the perceiver-resampler, which we denote with `(Perc)`.

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 the POPE metric/protocol). 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 C. We did not separately optimize hyperparameters for each LLM and will thus refrain from their mutual performance comparison; instead, for

[6]Without VQA in the training mix, the LVLMs do 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.
`Llama-3` Base	86.87	81.73	75.83	**83.13**	70.47	65.63	**78.53**	66.13	58.20
`Llama-3 +GC`	86.83	82.43	78.90	81.87	71.60	68.50	77.57	**67.70**	60.37
`Llama-3 +RE`	84.10	81.87	**79.93**	76.07	**73.10**	**71.73**	70.53	67.07	**64.57**
`Llama-3 +RE+GC`	**84.70**	**83.77**	**79.93**	75.47	71.00	69.73	67.63	64.50	61.27
`Phi-3` Base	87.17	85.30	81.87	81.57	77.57	73.73	**79.10**	**74.77**	66.40
`Phi-3 +GC`	85.30	83.73	81.80	78.93	75.53	73.47	72.43	69.50	65.80
`Phi-3 +RE`	86.43	**85.50**	**83.50**	78.93	76.20	**74.10**	75.17	72.40	**68.83**
`Phi-3 +RE+GC`	**87.57**	85.43	81.77	**84.63**	**78.27**	74.00	77.03	74.30	68.30
`Vicuna` Base	87.23	84.03	81.40	81.10	74.17	70.80	**78.80**	**74.53**	64.10
`Vicuna +GC`	85.73	83.93	81.43	83.17	76.20	73.17	73.57	69.27	65.73
`Vicuna +RE`	85.30	84.07	81.90	79.83	**76.40**	**74.67**	76.00	71.43	65.83
`Vicuna +RE+GC`	**88.27**	**86.10**	**82.37**	**84.37**	75.77	73.13	77.93	72.53	**65.80**
`Vicuna (Perc)` Base	**85.90**	**82.73**	78.00	**79.37**	69.40	65.10	**76.60**	67.27	57.80
`Vicuna (Perc) +GC`	83.93	82.23	78.33	76.37	69.77	64.97	73.20	66.47	59.20
`Vicuna (Perc) +RE`	83.63	82.60	**78.37**	76.40	**73.13**	**70.03**	69.13	**68.03**	**62.33**
`Vicuna (Perc) +RE+GC`	84.97	80.27	76.03	78.20	71.30	67.90	71.87	65.90	60.27

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	R+	Rg	R
`Llama-3 +RE`	60.02	53.69	65.41
`Llama-3 +RE+GC`	64.62	60.51	71.50
`Phi-3 +RE`	63.33	61.06	67.09
`Phi-3 +RE+GC`	68.23	65.50	73.33
`Vicuna +RE`	58.03	58.78	61.89
`Vicuna +RE+GC`	68.25	65.30	73.66
`Vicuna (Perc) +RE`	23.00	22.21	30.60
`Vicuna (Perc) +RE+GC`	35.68	34.32	42.20

Table 2: Precision@50 for expression grounding (provide the bounding box for a region) for the test split of RefCOCO (R), RefCOCO+ (R+), and RefCOCOg (Rg).

each of the three LLMs, we analyze how inclusion of grounding objectives affects their hallucination.

Referring Expressions. Before we test the effects of grounding on free-form and QA-based hallucination, we first analyze if the two grounding objectives are mutually compatible. Concretely, we test how the models trained with grounding objectives (`+RE`, and `+RE+GC`) perform on one of the grounding tasks itself. In other words, we test if and how well models explicitly trained with grounding objectives learn to ground expressions and whether the two grounding objectives are mutually beneficial. The results for expression grounding (one of the two `RE` tasks: given the description, provide the bounding box) are shown in Table 2. The metric is precision@50, that is, the proportion of examples where the intersection between the predicted and gold bounding box contains at least 50% of their union. The results indicate that adding grounded captioning (`+GC`) consistently and substantially improves the performance for all three LLMs: this strongly suggests that the two grounding objectives are mutually compatible. Vicuna-based model with the perceiver-resampler (`Perc`) aligner considerably underperforms the (default) MLP aligner; we suspect that this is because the (pre-)training data was insufficient for it to learn to properly encode positional information.

QA Hallucinations with POPE. Table 1 summarizes the hallucination results according to the QA-based evaluation protocol with POPE. Overall, both grounding objectives, referring expressions (`+RE`) and grounding captions (`+GC`) fail to consistently and non-negligibly improve performance, i.e., reduce hallucination. While their combination `+RE+GC` greatly improves grounding capabilities over `+RE` alone for all LLMs (Table 2), the same is not true for QA-based hallucination reduction (i.e., POPE), pointing to the lack of causal link between object grounding and hallucination reduction.

Standard Captions. Table 3 displays the performance of our LVLM variants on standard image captioning. We observe consistently, for all tested models on both evaluation datasets, that grounding objectives (i.e., their inclusion or exclusion) have little to no effect on performance: all models learn to generate proper captions in the MSCOCO style, with 10 words on average and of similar general quality, as captured by the caption quality metrics (CIDEr, CLIPScore). The metrics that capture caption detailness (coverage, number of objects & atomic facts) also show little difference between the models. Most importantly, the same is true for hallucination metrics CHAIR$_i$ and FaithScore,

	Model	CIDEr↑	CLIPS.↑	#Words	CHAIR$_i$ ↓	Coverage↑	Objects	FaithScore↑	Facts
MSCOCO	Llama-3 Base	112.31	11.71	10.22	3.84	56.43	1.61	91.25	4.49
	Llama-3 +GC	110.40	11.33	10.68	3.61	54.34	1.56	90.74	4.50
	Llama-3 +RE	109.01	11.36	10.52	3.78	55.74	1.60	90.86	4.64
	Llama-3 +RE+GC	107.95	11.72	10.66	3.63	55.46	1.61	90.64	4.69
	Phi-3 Base	112.54	11.97	11.41	3.28	57.54	1.68	90.98	4.88
	Phi-3 +GC	114.78	12.15	11.06	3.83	56.55	1.66	90.90	4.79
	Phi-3 +RE	113.22	12.07	11.14	3.43	57.18	1.68	91.06	4.87
	Phi-3 +RE+GC	113.68	11.90	11.06	3.68	56.21	1.64	91.28	4.66
	Vicuna Base	115.57	11.93	10.31	3.68	54.14	1.56	91.95	4.61
	Vicuna +GC	117.35	11.80	9.82	3.08	53.98	1.50	92.05	4.37
	Vicuna +RE	112.06	11.76	9.92	3.41	54.21	1.55	92.19	4.53
	Vicuna +RE+GC	113.30	11.77	9.79	3.64	52.69	1.50	91.98	4.27
	Vicuna (Perc) Base	107.74	11.27	10.05	4.73	53.71	1.55	90.56	4.46
	Vicuna (Perc) +GC	110.61	11.50	9.86	4.16	54.11	1.53	90.53	4.35
	Vicuna (Perc) +RE	107.38	11.31	9.96	4.54	54.21	1.57	90.66	4.51
	Vicuna (Perc) +RE+GC	109.64	11.25	10.11	5.15	54.20	1.57	90.39	4.56
Objects365	Llama-3 Base	—	10.99	10.15	14.51	27.67	1.94	88.68	4.56
	Llama-3 +GC	—	10.84	10.72	13.33	26.72	1.84	88.88	4.52
	Llama-3 +RE	—	10.67	10.50	12.74	26.73	1.86	88.57	4.66
	Llama-3 +RE+GC	—	10.98	10.74	12.48	28.16	1.96	87.97	4.86
	Phi-3 Base	—	11.27	11.36	12.99	29.23	2.03	88.33	4.77
	Phi-3 +GC	—	11.60	11.08	13.17	28.73	1.96	88.90	4.70
	Phi-3 +RE	—	11.41	11.22	13.30	28.20	1.97	89.06	4.88
	Phi-3 +RE+GC	—	11.31	11.18	12.27	28.78	1.97	88.93	4.64
	Vicuna Base	—	11.06	10.28	12.44	27.38	1.88	88.81	4.55
	Vicuna +GC	—	11.12	9.78	12.62	26.23	1.76	89.82	4.24
	Vicuna +RE	—	10.93	10.17	12.85	26.96	1.84	89.33	4.58
	Vicuna +RE+GC	—	11.07	9.83	12.60	26.25	1.79	90.20	4.24
	Vicuna (Perc.) Base	—	10.14	10.12	15.82	25.82	1.87	86.18	4.36
	Vicuna (Perc) +GC	—	10.52	9.81	14.42	25.50	1.74	87.65	4.19
	Vicuna (Perc) +RE	—	10.24	10.26	15.81	25.98	1.88	86.07	4.55
	Vicuna (Perc) +RE+GC	—	10.30	10.23	16.68	25.92	1.84	86.50	4.48

Table 3: Results on standard image captioning. CIDEr and CLIPScore indicate general 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.

confirming that there is **no** positive transfer from grounding to hallucination reduction.

Grounded Captions. Previous results establish that *training* on grounding objectives does not reduce hallucination in open caption generation. We next test whether forcing the model to generate grounded captions at *inference* can reduce hallucination. Intuitively, prompting the model to produce grounded captions should encourage it to generate only objects contained in the image. The results in Table 4 show that generating grounded captions indeed results in some hallucination reduction, but the effect is rather small. Reduction is more prominent on Objects365 where the baseline hallucination rate is higher than on MSCOCO. On the flip side, generating grounded captions at inference slightly reduces their informativeness too (i.e., we observe fewer objects and atomic facts in the generated captions). A closer qualitative inspection (see §6) reveals that LVLMs trained with grounding objectives still incorrectly describe objects or fabricate them entirely.

6 Qualitative Grounded Caption Analysis

We show examples for grounded captioning in Figure 2. The grounding itself does not necessarily prevent the model from hallucinating: in the first example, the model fully hallucinates a woman along with a bounding box for her. In the second example, the second 'elephant' bounding box is positionally correct in that it points to an animal, but that animal is a rhino. In the third example, similarly, the bounding box correctly contains an apple but the attribute 'peeled' is hallucinated. These

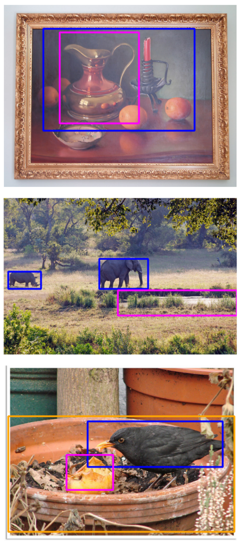

Figure 2: Qualitative examples of Vicuna +RE+GC for standard and grounded captioning. Hallucinations are underlined in red. Predicted bounding boxes are visualized in the image and marked in the caption.

	Model	CIDEr↑	CLIPS.↑	#Words	CHAIR$_i$ ↓	Coverage↑	Objects	FaithScore↑	Facts
MSCOCO	Llama-3 +GC	-8.52	0.28	-0.48	0.17	-5.63	-0.21	1.12	-0.18
	Llama-3 +RE+GC	-7.92	-0.20	-0.44	-0.39	-5.44	-0.25	0.88	-0.28
	Phi-3 +GC	-6.23	-0.25	-0.34	-0.14	-6.33	-0.28	0.63	-0.41
	Phi-3 +RE+GC	-8.12	-0.17	-0.24	0.44	-7.36	-0.28	1.08	-0.29
	Vicuna +GC	-9.32	-0.03	0.46	0.51	-6.64	-0.19	0.72	-0.09
	Vicuna +RE+GC	-8.22	0.09	0.91	0.03	-4.80	-0.19	0.48	0.11
	Vicuna (Perc.) +GC	-7.78	-0.22	0.12	0.06	-6.87	-0.22	0.61	-0.18
	Vicuna (Perc.) +RE+GC	-13.69	-0.16	0.23	-1.08	-8.13	-0.32	0.87	-0.19
Objects365	Llama-3 +GC	—	-0.02	-0.50	-1.07	-3.06	-0.25	0.46	-0.18
	Llama-3 +RE+GC	—	-0.34	-0.31	-0.01	-3.67	-0.32	1.09	-0.30
	Phi-3 +GC	—	-0.39	-0.03	-1.91	-2.89	-0.26	0.87	-0.22
	Phi-3 +RE+GC	—	-0.28	-0.05	-0.48	-3.12	-0.28	0.74	-0.09
	Vicuna +GC	—	0.04	0.44	-1.38	-2.03	-0.17	0.21	0.09
	Vicuna +RE+GC	—	-0.06	0.86	-1.06	-3.35	-0.27	-0.25	0.26
	Vicuna (Perc.) +GC	—	-0.00	0.25	-0.77	-2.61	-0.21	-0.14	0.03
	Vicuna (Perc.) +RE+GC	—	-0.12	0.30	-2.37	-3.40	-0.37	1.59	-0.06

Table 4: Absolute performance difference of grounded image captioning w.r.t. standard captioning (Table 3).

examples point to causes of hallucination that go beyond insufficient or incorrect grounding and help explain why grounding objectives do not really reduce the LVLM hallucination in open captioning.

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 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 hallucination remains one of the main obstacles to wide-range adoption of LVLMs. Prior work suggested that grounding objectives like referring expressions reduce hallucination but the empirical support for this claim is confined to QA-based evaluation. In this work, we carried out an in-depth analysis of the effects that grounding objectives in LVLM training have on their hallucination in open image captioning. Our extensive experiments with three backbone LLMs show that there is *no* causal link between improved object grounding (via objectives like referring expressions) and hallucination reduction: this observation is true both under QA-based and open captioning hallucination evaluation protocols. Finally, we observe that explicitly prompting LVLMs to generate grounded captions at inference can slightly reduce hallucination but at the expense of reduced caption informativeness.

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, 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, additional or larger LLMs, and/or alignment modules other than the MLP or 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

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Parul Chopra, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Dan Iter, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Chen Liang, Weishung Liu, Eric Lin, Zeqi Lin, Piyush Madan, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Xia Song, Masahiro Tanaka, Xin Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Michael Wyatt, Can Xu, Jiahang Xu, Sonali Yadav, Fan Yang, Ziyi Yang, Donghan Yu, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. _eprint: 2404.14219.

AI@Meta. 2024. Llama 3 Model Card.

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.

Xiangxiang Chu, Limeng Qiao, Xinyu Zhang, Shuang Xu, Fei Wei, Yang Yang, Xiaofei Sun, Yiming Hu, Xinyang Lin, Bo Zhang, and Chunhua Shen. 2024. MobileVLM V2: Faster and Stronger Baseline for Vision Language Model. *CoRR*, abs/2402.03766. ArXiv: 2402.03766.

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.

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.

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. 2021a. 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.

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. 2021b. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint*, abs/2103.00020. _eprint: 2103.00020.

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.

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.

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.

Model	#Words	$\text{CHAIR}_i \downarrow$	Coverage↑	Objects
`Llama-3 Base`	94.46	30.78	44.45	7.44
`Llama-3 +GC`	100.61	31.74	44.80	8.08
`Llama-3 +RE`	100.39	29.08	43.66	7.57
`Llama-3 +RE+GC`	103.75	26.42	43.86	7.66
`Phi-3 Base`	99.17	27.18	46.16	7.00
`Phi-3 +GC`	94.33	25.69	45.45	6.97
`Phi-3 +RE`	97.09	27.75	45.20	6.85
`Phi-3 +RE+GC`	96.55	27.74	45.69	7.12
`Vicuna Base`	93.91	26.10	45.12	7.18
`Vicuna +GC`	89.69	25.61	44.42	7.25
`Vicuna +RE`	96.45	28.76	43.20	6.94
`Vicuna +RE+GC`	90.18	26.06	44.10	7.28
`Vicuna (Perc.) Base`	93.98	31.52	41.18	7.02
`Vicuna (Perc.) +GC`	92.64	31.28	40.67	7.24
`Vicuna (Perc.) +RE`	96.39	32.79	40.15	7.08
`Vicuna (Perc.) +RE+GC`	96.14	35.10	41.32	7.94

Table 6: Results for long captions on Objects365. We report the average number of words and CHAIR metrics. Results with FaithScore and on MSCOCO are qualitatively the same so we omit them for brevity.

A Training and Details

All models were trained on a single NVIDIA RTX3090s card, with training duration ranging between 2-4 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) and a cosine schedule. For LoRA, we set $r = 64, \alpha = 128$. During pre-training, where only the parameters of the alignment module are updated, we use batch size 32, learning rate 0.001, and weight decay 0. For training on the task mix, we use learning rate 2e-4, weight decay 0, and batch size 16/32/64 for Vicuna/Phi-3/Llama-3 (achieved with gradient accumulation).

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., $[0.10, 0.05, 0.64, 1.00; 0.50, 0.15, 0.64, 1.00]$). If we would have more than three boxes in brackets, we instead create a single bounding box covering all boxes to limit the final sequence length.

B Long Captions

Table 6 shows long captioning results. For brevity, we only report the results for Objects365 with CHAIR(-MEN): for MSCOCO and FaithScore the results are qualitatively the same. Overall, the differences between model variants are negligible similar to the standard captions. The grounding objectives (`+RE` and `+GC`) thus does not seem to affect long captions. This again questions the extent to which improved fine-grained image understanding from grounding actually transfers to hallucination reduction in open generation.

C 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 7, we report string-matching CHAIR results for MSCOCO, which can be compared to Table 3 (standard captions), Table 4 (grounded captions), and Table 6 (long captions).

We find that results with CHAIR-MEN are highly proportional to CHAIR. This validates CHAIR-MEN as an alternative approach for identifying hallucinated objects and opens up the extension to other datasets like Objects365.

Model	CHAIR$_i$ ↓	Coverage↑	Objects
Llama-3 Base	4.36	58.84	1.62
Llama-3 +GC	4.12	57.30	1.57
Llama-3 +RE	4.36	58.06	1.61
Llama-3 +RE+GC	5.30	59.41	1.68
Phi-3 Base	4.26	60.39	1.70
Phi-3 +GC	4.39	59.79	1.67
Phi-3 +RE	4.41	59.73	1.69
Phi-3 +RE+GC	4.44	59.21	1.67
Vicuna Base	4.45	58.62	1.62
Vicuna +GC	3.46	57.74	1.55
Vicuna +RE	4.14	57.78	1.59
Vicuna +RE+GC	3.92	56.80	1.55
Vicuna (Perc.) Base	5.66	57.50	1.60
Vicuna (Perc.) +GC	4.87	57.10	1.55
Vicuna (Perc.) +RE	5.38	57.57	1.60
Vicuna (Perc.) +RE+GC	6.08	58.33	1.62

(a) MSCOCO Standard Captions

Model	CHAIR$_i$ ↓	Coverage↑	Objects
Llama-3 +GC	4.32	53.21	1.41
Llama-3 +RE+GC	5.21	54.71	1.48
Phi-3 +GC	4.03	54.61	1.44
Phi-3 +RE+GC	3.49	54.28	1.43
Vicuna +GC	3.98	52.66	1.38
Vicuna +RE+GC	3.33	53.54	1.41
Vicuna (Perc.) +GC	4.78	52.29	1.38
Vicuna (Perc.) +RE+GC	6.65	52.37	1.41

(b) MSCOCO Grounded Captions

Model	CHAIR$_i$ ↓	Coverage↑	Objects
Llama-3 Base	23.45	80.62	7.10
Llama-3 +GC	24.54	80.02	7.62
Llama-3 +RE	23.22	79.37	7.55
Llama-3 +RE+GC	20.63	79.23	7.20
Phi-3 Base	20.92	81.05	6.28
Phi-3 +GC	18.10	78.89	6.13
Phi-3 +RE	21.01	79.32	5.82
Phi-3 +RE+GC	22.16	79.82	6.31
Vicuna Base	17.54	80.17	6.51
Vicuna +GC	17.70	78.76	6.33
Vicuna +RE	18.27	79.59	6.16
Vicuna +RE+GC	18.20	78.68	6.49
Vicuna (Perc.) Base	23.35	77.82	6.71
Vicuna (Perc.) +GC	22.19	77.11	6.76
Vicuna (Perc.) +RE	22.74	77.85	6.67
Vicuna (Perc.) +RE+GC	24.83	78.09	7.31

(c) MSCOCO Long Captions

Table 7: CHAIR results for MSCOCO using the classic string-matching approach.