THRONE: An Object-based Hallucination Benchmark for the Free-form Generations of Large Vision-Language Models

Prannay Kaul[1*] Zhizhong Li[2†] Hao Yang [2] Yonatan Dukler [2]
Ashwin Swaminathan [2] C. J. Taylor [2] Stefano Soatto [2]
VGG, University of Oxford[1] AWS AI Labs[2]
prannay@robots.ox.ac.uk {lzhizhon,haoyng,dukler,swashwin,taylorcj,soattos}@amazon.com

Abstract

Mitigating hallucinations in large vision-language models (LVLMs) remains an open problem. Recent benchmarks do not address hallucinations in open-ended free-form responses, which we term "Type I hallucinations". Instead, they focus on hallucinations responding to very specific question formats—typically a multiple-choice response regarding a particular object or attribute—which we term "Type II hallucinations". Additionally, such benchmarks often require external API calls to models which are subject to change. In practice, we observe that a reduction in Type II hallucinations does not lead to a reduction in Type I hallucinations but rather that the two forms of hallucinations are often anti-correlated. To address this, we propose THRONE*, a novel object-based automatic framework for quantitatively evaluating Type I hallucinations in LVLM free-form outputs. We use public language models (LMs) to identify hallucinations in LVLM responses and compute informative metrics. By evaluating a large selection of recent LVLMs using public datasets, we show that an improvement in existing metrics do not lead to a reduction in Type I hallucinations, and that established benchmarks for measuring Type I hallucinations are incomplete. Finally, we provide a simple and effective data augmentation method to reduce Type I and Type II hallucinations as a strong baseline.*

1 Introduction

This paper proposes a benchmark to evaluate hallucinations by large vision-language models (LVLMs) when generating free-form responses, specifically detailed descriptions, based on a given image.

The rapid advancement in large language models (LLMs) [52] has pushed the development of large vision-language models (LVLMs). [1, 6, 7, 17, 24, 25, 30, 36, 44, 48, 54] LVLMs take input text *and images* and generate text responses to enable multi-modal perception and comprehension.

LVLMs are largely built on LLMs and therefore inherit both their advantages and their disadvantages. LLMs have been shown to produce hallucinations [47, 51], generated text responses that are coherent and plausible but factually incorrect. LVLMs echo this behavior with generated text contradicting with the visual or text input [53]. Hallucinations prevent the use of LVLMs in safety-critical situations and therefore evaluating and mitigating hallucinations in LVLMs is crucial for their deployment in such settings. Determining the presence and cause of hallucinations in LVLMs remains an open question [45, 53].

We divide LVLM hallucinations into two types. Type I hallucinations occur in response to open-ended questions with a very large set of possible responses —*e.g.* `What is happening in this image?`. Type II hallucinations are incorrect responses to a factual question regarding a specific concept about the image with a fixed set of options such as yes/no—*e.g.* `Is there a traffic light in this image?` Fig. 3 illustrates the difference between these two types of hallucination. Reducing hallucinations in both cases is required for useful, multi-purpose LVLMs. However, later in Fig. 4 we observe that the same LVLM can give contradicting answers when being evaluated for Type I vs. Type II hallucinations. This implies that measuring and reducing one type does not necessarily reduce the other.

Existing works to evaluate LVLMs often avoid direct quantification of hallucinations and instead develop comprehensive benchmarks that judge various other desirable abilities such as: optical character recognition, fine-grained recognition and attribute detection [14, 22, 32]. The extent of hallucinations in these benchmarks is obfuscated, since it is only one of many factors influencing other metrics. It requires human effort to inspect individual predictions. There are two major established works which specifically develop a benchmark for evaluating hallucinations in vision-language models: POPE [26] and CHAIR [40], which we discuss in detail in Sec. 2. However, we observe they both have shortcomings in effectively evaluating hallucinations:

[*] Work conducted during an internship at Amazon
[†] Corresponding author

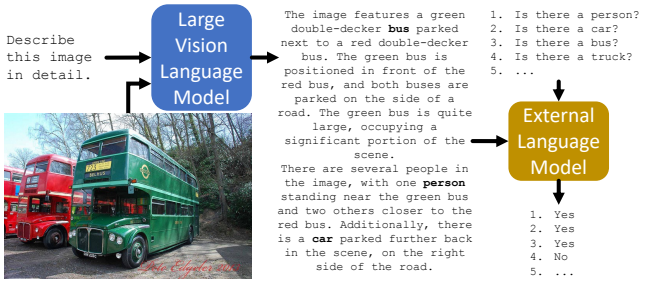

Figure 1. **THRONE (Ours):** LVLMs are prompted with a concept neutral instruction. An external LM performs abstractive QA on the response to establish the existence of **Type I** hallucinations.

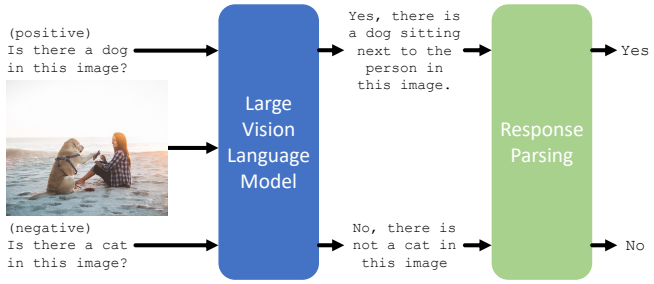

Figure 2. **POPE:** Questions with specific concepts prompt an LVLM directly to evaluate **Type II** hallucinations [26]. Hand-crafted rules parse LVLM responses to give yes/no labels.

POPE [26] is a recent work addressing Type II hallucinations with respect to object classes. However, we find Type I and Type II hallucinations are disconnected, and that POPE gives an incomplete picture on LVLM hallucinations. Moreover, POPE systematically under-samples negative object categories leading to a large underestimation of Type II hallucinations (see Sec. 5.4).

CHAIR [40] does address Type I hallucinations—establishing object category hallucination in short image captions using simple text matching. However, CHAIR is not suited to current LVLMs because the simple text matching it employs cannot comprehend abstract or hypothetical concepts present in today's LVLMs (see Fig. 4). Further, hand-crafted rules for each set of classes are required for usable text matching, and trivial model answers can attain a perfect CHAIR score.

To address the issues of current LVLM hallucination benchmarks, we propose THRONE (***T**ext-from-image **H**allucination **R**ecognition with **O**bject-probes for open-e**N**ded **E**valuation*). THRONE leverages language models (LMs) to evaluate Type I hallucinations in free-form, open-ended image descriptions with respect to a pre-defined object vocabulary of interest. By utilizing LMs, THRONE is able to accurately judge whether an object mentioned in an LVLM response is implied to exist in the image or is abstractly mentioned with no implication about its existence (see Fig. 4).

Moreover, in THRONE, we provide easy access to our benchmark, by leveraging open-source LMs that can run on common GPUs, instead of relying on closed-source commercial models [35] that are subject to arbitrary change, as done in other works [4, 22, 32]. Through combining multiple open-source LMs, we mitigate any single-model biases in judging hallucinations when calculating Type-I hallucination scores with THRONE.

We make four contributions: *first*, we establish an accurate and accessible benchmark to quantitatively evaluate object hallucinations in free-form responses, leveraging LMs to judge the existence of Type I hallucinations—quantitatively showing half the judgement errors of CHAIR; *second*, we evaluate a number of current LVLMs on THRONE and demonstrate that observed progress in reducing Type II hallucinations do not translate to a corresponding reduction in Type I hallucinations; *third*, we show a recent method, POPE, is inaccurately capturing the extent of Type II hallucination discovery due to its sampling strategy; we mitigate this issue in our implementation of THRONE while presenting results for a complete version of POPE; and *fourth*, we provide a simple augmentation of visual instruction tuning data which significantly improves performance on Type I hallucinations while maintaining or improving Type II hallucination performance.

2 Related Work

Hallucination Benchmarks for LVLMs. In response to the development of LVLMs, few evaluation benchmarks focusing on hallucinations have been introduced. CHAIR [40], one of the first works to assess hallucinations, is designed for short *image captions*. CHAIR uses a fixed set of object classes (extended with their synonyms) as a set of text strings to find predicted object classes in image captions via exact text matching. Subsequently, each class matched in the caption is compared to the COCO [27] captions and bounding box annotations to establish object hallucinations. CHAIR was developed prior to current instruction-tuned LVLMs which produce long free-form responses (10× longer than image captions) with a diverse vocabulary, limiting its applicability to modern LVLMs. As shown in Fig. 3, exact text matching in CHAIR is prone to incorrectly matching the vocabiliary classes with abstract concepts in the free-form response and the and synonyms of the class list must be manually selected to prevent confusion during evaluation. For example, in CHAIR the word "`chair`" will match all responses for the phrase "`toilet seat`", because the exact text matching classifies "`seat`" as the COCO class "`chair`". Finally, CHAIR metrics compare the overall number of predicted objects to the overall number of predicted objects judged to be hallucinations—ignoring the recall of ground-truth objects and the distribution of object classes. This means that a single correct

prediction across the entire evaluation dataset along with a generic response otherwise *e.g.* `A natural scene`, achieves a perfect score ($\frac{\text{0 hallucinated objects}}{\text{1 predicted objects}} = 0.0$). See the Supplementary Material for a full overview of CHAIR. In contrast, THRONE uses pre-trained LMs that go beyond direct synonym matching, to automatically judge the existence of concepts and hallucinations in free-form responses. In addition, our method considers both recall and precision to yield a holistic benchmark and does not require any manual curation of synonyms. THRONE and CHAIR both evaluate Type I hallucinations—hallucinations in response to concept-neutral prompts *e.g.* `Describe this image in detail.`

POPE [26] is a recently proposed benchmark to evaluate object hallucinations in LVLMs—specifically Type II hallucinations, in which an LVLM is directly queried with a yes-no question regarding the existence of a particular object of the form: `Is there {a/an} {object_class_name} in the image?`. The LVLM response is parsed using simple rules to determine whether a Type II hallucination has occurred. Precision and recall metrics are compiled using the parsed LVLM responses and the ground-truth annotation data. Despite focusing on measuring hallucinations, POPE only queries an LVLM with 3 positive and 3 sampled negative questions per COCO image *i.e.* the evaluation is artificially balanced. This means many potential hallucinations with respect to the COCO categories are not captured by their method. In Sec. 5.4, we show POPE dramatically underestimates Type II hallucinations and present the results of a complete version. Our method, THRONE, evaluates the prevalence of Type I hallucinations, which we observe are disjoint to Type II hallucinations.

Comprehensive Benchmarks for LVLMs have recently grown in number. MMBench [32] and MM-Vet [49] assess various aspects of LVLM performance such as: color perception, celebrity recognition, and numerical calculation. However many of these works integrate evolving APIs which are modified often (or even discontinued) and are inherently stochastic. Over time this greatly reduces the consistency of these benchmarks. Exceptions are MME [14] and SEED-Bench [22], but the impact of Type II hallucinations on final metrics is conflated with a number of other aspects of model performance. Our method, THRONE, directly addresses Type I hallucinations, only making use of open-source language models and datasets.

Large Vision Language Models (LVLMs) have rapidly developed by harnessing advancements in large language models (LLM) [5, 39, 43] and by directly integrating pretrained LLMs into their architectures. In contrast to earlier vision-language models such as CLIP [18, 38], LVLMs are generally comprised of a pretrained LLM and image encoder, aligned with a connector module of varying complexity. Some works are highlighted here. Frozen [44] is an early work fine-tuning the vision encoder to dynamically prefix the prompt to a frozen LLM. Flamingo [1] combines visual and language features using cross-attention layers in an otherwise frozen LLM. BLIP-2 [25] uses a frozen image encoder to learn a Querying-Transformer (Q-Former) on image-text pairs that is used as the connector. This architecture is adapted in [7] for dialogue via training with visual instruction tuning. LLaVA [30] uses COCO [27] annotations and GPT-4 [35] to generate visual instruction tuning data in a plain-text pipeline. Combining this generated data with standard VQA datasets (VQAv2 etc.) further boosts performance [29]. Different works modify the training approach by using efficient adaptation [16, 48] multiple training stages [3] or introduce the use of discrete tokens for localization [36]. We note however, that most of these works evaluate performance on traditional vision-language datasets like VQAv2 [2], which do not consider the extent of hallucinations, a known problem with LLMs [47, 51].

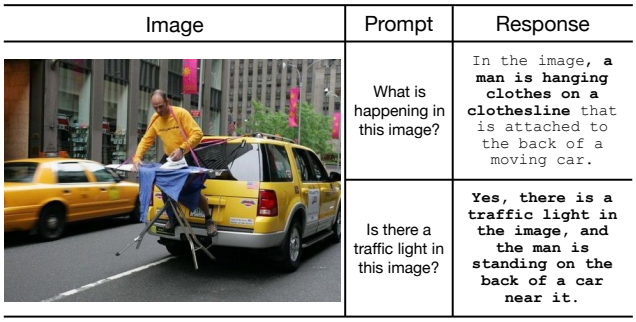

Figure 3. **Type I vs. Type II Hallucinations:** (Top) LVLMs prompted with concept-neutral instructions produce Type I hallucinations. (Bottom) Instructions specifying a concept produce Type II hallucinations. Examples from LLaVA-v1.5 [29].

3 THRONE

Recap of existing methods. POPE [26] and other benchmarks (MME [14], MMBench [32]) directly query LVLMs with a restricted desired answer space, *e.g.* yes-no (MME, POPE) and A-B-C-D multiple choice (MMBench), as shown in Fig. 2. These benchmarks only consider such short answer formats, whereas a key quality of LVLMs is in their ability to generate free-form coherent text. Moreover, POPE, which addresses Type II hallucinations, undersamples negative classes meaning hallucinations are dramatically underestimated (see Sec. 5.4 and Fig. 7). In contrast, we skip class subsampling and *enumerate all classes for every image*, ensuring a full evaluation of Type I hallucinations of the ground-truth classes.

CHAIR [40] also evaluates Type I hallucinations, but was developed when typical vision-language models could only generate short and simple captions similar to those in COCO Captions [12]. Moreover, it lacks accurate comprehension of responses (see the right side of Fig. 4) and ig-

Type II Hallucination Evaluation (POPE)

Instruction [GT]	Response [Prediction]
Is there...	
a person [No]	No, there is no person in the image. [No]
a banana [Yes]	Yes, there is a banana in the image. [Yes]
an apple [No]	No, there is not an apple in the image. [No]
an orange [No]	No, there is no orange in the image. [No]
...in the image?	***No Type II Hallucinations Found***

Type I Hallucination Evaluation

Instruction	Response
Describe this image in detail. - GT Class - Type II Hallucination - Hypothetical Content (not a Hallucination)	The image showcases a fruit stand at a grocery store, featuring a variety of fruits on display. There are several bunches of **bananas**, with some placed in the foreground and others in the background. The **bananas** are arranged in different sections, creating an appealing presentation for **customers**. In addition to the **bananas**, there are also **apples** and **oranges** on display. The **apples** are located towards the left side of the image, while the **oranges** are placed in the middle and right side of the stand. The fruits are well-organized and presented in an attractive manner, making it an inviting sight for **shoppers**. ***Type I Hallucinations Present and Found***

MSCOCO Object Prediction *from Description*

Human	CHAIR	THRONE (Ours)
banana **apple** **orange**	**banana** **person** **apple** **orange**	**banana** **apple** **orange**

Figure 4. **A Comparison of POPE, CHAIR and THRONE:** Directly querying LVLMs for object existence (person, banana etc.) using concept-specific instructions, as in POPE (bottom left), does not produce the same hallucinations as using concept-neutral instructions (right). We highlight the Type I hallucinations in orange. CHAIR relies on exact text matching to a fixed set of objects and synonyms, thus incorrectly labels "customers" and "shoppers" as hallucinations, highlighted in red. THRONE is designed for the rich vocabulary and the free-form generations of modern LVLMs by harnessing LMs to establish object existence. By using an LM to pass judgement, our evaluation correctly captures "customers" and "shoppers" as hypothetical content in the free-form generation.

nores the recall of ground-truth objects. In Sec. 5.5, we describe quantitative evaluations, using a human oracle, which demonstrate THRONE *halves* the rate of hallucination misjudgement in CHAIR. See Sec. 2 and the Supplementary Material for more details on CHAIR and its shortcomings. Fig. 4 shows an overview of the three aforementioned methods: POPE, CHAIR and our method, THRONE.

3.1 Evaluating Hallucinations with THRONE

To address these limitations, we propose a framework, THRONE, shown in Fig. 1, to evaluate the prevalence of Type I hallucinations in LVLM responses conditioned on an image and a neutral text prompt.

For each image in a labeled dataset, $\mathcal{I}$, addressing a set of classes, $\mathcal{C}$, the LVLM is queried with the same instruction: `Describe this image in detail.`, regardless of image content. The LVLM response, which is expected to be long free-form text containing an image description, is generated and stored. Next, a publicly available, open-source, *external* language model (LM) performs *abstractive question answering* (AQA) using the LVLM response as context and a question of the form: `Please answer yes or no. Is there {a/an} {object class name} in this image?` or similar, *for every class* in $\mathcal{C}$ (right side of Fig. 1). By selecting an appropriate LM and using a simple prompt template (see Sec. 4 for specific details), we ensure the AQA response is either `yes` or `no`—our method does not require any additional parsing. This is in contrast to other works which require added parsing or interpretation by a closed-source model.

After performing AQA on each response generated by the LVLM for every class in $\mathcal{C}$, we obtain an array of predicted labels:

$$\hat{\mathbf{Y}} \in \{0,1\}^{|\mathcal{I}|\times|\mathcal{C}|} \tag{1}$$

where 0/1 indicates a negative/positive existence *judgement* by the LM with respect to the relevant LVLM response.

Similarly using the ground-truth data for $\mathcal{I}$, an array of ground-truth labels can be constructed:

$$\mathbf{Y} \in \{0,1\}^{|\mathcal{I}|\times|\mathcal{C}|} \tag{2}$$

Using these two arrays, we calculate four metrics: (1) Overall Precision, P_{ALL}; (2) Overall Recall, R_{ALL}; (3) Class-wise Precision, P_{CLS}; and (4) Class-wise Recall, R_{CLS}. Overall metrics are calculated in a class-agnostic manner. Class-wise metrics are calculated in a class-conscious manner by computing precision and recall for each category separately and then averaging. This follows common practice in object detection and instance segmentation [10, 27].

False positives in LVLMs reduce precision and are dominated by hallucinations—precision indicates the extent of Type I hallucinations in LVLM responses. The recall metrics inform the level of class coverage by an LVLM when producing image descriptions. The class-wise metrics give a general measure of performance as the overall metrics are skewed towards the most common categories. A common way to combine precision, P, and recall, R, metrics is through the generalized F score, F_β:

$$F^\beta = (1+\beta^2)\cdot\frac{P\cdot R}{(\beta^2\cdot P)+R} \tag{3}$$

Prior work such as POPE [26], use the common balanced F-score (or F^1-score) which equally weights precision and

recall. However, given that THRONE is concerned with measuring hallucinations and is therefore particularly interested in precision, we choose $\beta = 0.5$ or the $F^{0.5}$-score, thus weighing precision twice as important as recall. The $F^{0.5}$-score is commonly used in pandemic misinformation filters [37], recommender systems [34], active stock selection [41] and other areas where false positives are costlier than false negatives. From this we can calculate the overall and class-wise $F^{0.5}$-scores, $F^{0.5}_{\text{ALL}}$ and $F^{0.5}_{\text{CLS}}$, respectively. To mitigate against class imbalance issues, we use $F^{0.5}_{\text{CLS}}$ as the principle metric of comparison between LVLMs in THRONE.

3.2 Ensuring Robustness via Ensembling

Any LM used for AQA in THRONE may misjudge Type II hallucinations—no LM is perfect. Fig. 5 (top) shows two different variants from the same model family (FLAN-T5 [33]), yielding opposite responses when prompted with the same response and question. Moreover, an LM may yield different answers to semantically identical questions, despite conditioning on the same response—shown in Fig. 5 (bottom). To ensure THRONE is robust to spurious performance by any one LM, we ensemble various LMs and semantically equivalent question formats. The use of N distinct LMs and M distinct question formats yields NM answers for each (LVLM response, class) combination. This set of NM answers is combined based on equal voting by each (LM, question) pair to "elect" the predicted answer. Continuing the notation from Eq. (1), stacking predictions from each (LM, question) pair yields a 3D array of predicted labels: $\bar{\mathbf{Y}} \in \{0,1\}^{|\mathcal{I}_{\text{OURS}}| \times |\mathcal{C}_{\text{OURS}}| \times NM}$. To combine the answers from each (LM, question) pair via voting, we require agreement between at least k answers. Where sufficient agreement does not exist we introduce an "ignore" label. Mathematically, the elements in $\hat{\mathbf{Y}}$ (from Eq. (1)) are calculated as:

$$\hat{Y}_{i,j} = \begin{cases} 0, & \sum_{k=1}^{NM} \bar{Y}_{i,j,k} \leq (NM - k) \\ 1, & \sum_{k=1}^{NM} \bar{Y}_{i,j,k} \geq k \\ -1, & \text{otherwise} \end{cases}$$

where an "ignore" label exists in $\bar{\mathbf{Y}}$, it is removed from the calculations of P_{ALL}, R_{ALL}, P_{CLS}, and R_{CLS}, as we cannot be confident in the AQA process for that particular (LVLM response, class) combination. The choice of k reflects the desired level of confidence. We make use of a *unanimous* voting mechanism ($k = NM$)—use the prediction only if *all* (LM, question) pairs agree, otherwise ignore. Human evaluation of our benchmark and choice of voting mechanism is found in the Supplementary Material. Once we have applied this voting mechanism, we can calculate the metrics described at the end of Sec. 3.1.

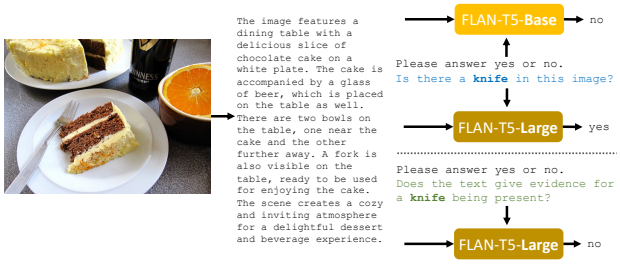

Figure 5. **AQA Ensembling in Evaluation:** Using different LMs or different prompts when running AQA on LVLM generated responses can produce opposing answers to *identical prompts* or *identical LMs*. To ensure THRONE is robust to this, we ensemble *multiple LMs* and *multiple prompts* in our evaluation pipeline.

4 Implementation

We provide details of our framework regarding the selection of public datasets, public LMs and LM prompts.

4.1 Dataset

Any proper evaluation of object hallucinations (a type of false positive error) requires knowing, with certainty, which classes are absent in an image. In our benchmark, we use COCO [27] for a number of reasons: (1) its annotations of 80 categories are exhaustive *at an image level* (image-level recall≈99% [27])—if there are many `book` instances in a COCO image, at least one is annotated with bounding boxes; (2) many LVLMs are partly trained on COCO data and so should be familiar with the set of categories; (3) its images generally contain complex scenes suitable for generating long free-form descriptions unlike image recognition datasets like ImageNet [8].

We utilize the validation set of COCO 2017, which contains $|\mathcal{I}| = 5000$ images and $|\mathcal{C}| = 80$ categories. Using the single LVLM text prompt `Describe this image in detail.`, we generate 5000 responses. As we query each LVLM response for each category in $\mathcal{C}$, a single LM performs AQA $|\mathcal{I}| \times |\mathcal{C}| = 400$k times across the LVLM responses—one instance of AQA per (image response, class) pair. In Sec. 5.3, we also present results using Objects365 [42] which is rarely used in LVLM training.

4.2 Language Models

To assess Type I hallucinations in an LVLM response using THRONE, we require a language model (LM) which can answer questions on the existence of object categories based on the LVLM response. MMBench [32] makes use of a ChatGPT model to identify multiple choice answer selections and still reports mistakes. In our experience, some LMs give rather incoherent judgements when used to assess hallucinations when the prompt is changed (see supplemental material). Therefore for THRONE, we choose FLAN-T5 models [33, 39]. We make this choice because FLAN-T5 model family: (1) have undergone instruction tuning with thousands of tasks [33]; (2) are open-source and

Model	L	P_{ALL}	R_{ALL}	F^1_{ALL}	$F^{0.5}_{\text{ALL}}$	P_{CLS}	R_{CLS}	F^1_{CLS}	$F^{0.5}_{\text{CLS}}$
Adapter-v2 [15]	514	63.6	73.3	68.1	65.3	68.2	70.6	69.4	68.7
Adapter-v2.1 [15]	512	63.8	73.7	68.4	65.5	67.4	71.2	69.3	68.1
InstructBLIP [7]	525	70.8	*74.3*	*72.5*	71.5	77.2	*71.9*	*74.5*	76.1
Otter-Image [23]	257	33.0	31.2	32.1	32.7	25.2	16.9	20.2	22.9
MiniGPT4 [54]	473	*81.7*	59.8	69.0	76.1	*79.9*	61.8	69.7	75.5
MiniGPT-v2 [6]	381	79.0	66.6	72.3	76.2	77.6	67.0	71.9	75.2
mPLUG-Owl [48]	555	55.5	71.9	62.6	58.1	66.3	68.3	67.3	66.7
LRV-Instruction-v2 [28]	103	*82.0*	56.7	67.0	75.3	*78.4*	58.8	67.2	73.5
LLaVA-v1.3* [30]	532	80.5	65.2	72.1	76.9	*79.9*	65.3	71.9	*76.5*
LLaVA-v1.5 [29]	509	68.1	61.0	64.4	66.6	69.9	56.4	62.5	66.8
LLaVA-Mistral [19, 31]	524	**86.8**	**71.8**	**78.3**	**83.6**	**84.4**	**64.2**	**70.8**	**77.5**

Table 1. **THRONE Results with COCO** for a selection of instruction-tuned LVLMs. We select $F^{0.5}_{\text{CLS}}$ as the principal metric for evaluation in our benchmark to balance across classes and to prioritize precision (which reflects the extent of hallucination) over recall. Best and second-best performance are denoted by **blue** and *red*, respectively. *Our implementation using official code to enable fair comparison. L corresponds to the median response length (measured in # of characters).

therefore accessible to the community; (3) can fit locally on a single GPU for the models we consider; (4) follow user's instruction to only respond `yes` or `no`; and (5) are optimized for use in the free and public Text-Generation-Inference API [11] for acceleration. As described in Sec. 3.2, we utilize N LMs to ensure our method is robust. Specifically, we use $N = 3$ variants of FLAN-T5, namely: FLAN-T5-Base (250M parameters), FLAN-T5-Large (780M parameters), and FLAN-T5-XL (3B parameters).

4.3 Prompt Ensembling

To guarantee each of these FLAN-T5 variants faithfully produce responses of either `yes` or `no` only during AQA, we use the following input template to each LM, reflecting the format used when training FLAN-T5[1]:

```
Text: {LVLM Response} Read the text
about an image and answer the question.
Question: Please answer yes or no.
{Question}
```

We use $M = 3$ semantically identical questions:

- `Is there a/an {class_name} in this image?`
- `Does the text imply a/an {class_name} is in the image?`
- `Does the text explicitly mention a/an {class_name} is in the image?`

As outlined in Sec. 3.2, we use a *unanimous voting* mechanism to combine the answers from each (LM, question) pair and so $k = 3 \times 3 = 9$.

5 Evaluation Results

In this section, we: (1) outline our LVLM selection and reasoning; (2) present THRONE on COCO for evaluating Type I hallucinations; (3) extend THRONE to Objects365 (containing a larger vocabulary); (4) analyze and extend POPE to enable improved evaluation of Type II hallucinations; and (5) highlight results from our ablation studies found in the Supplemental Material.

[1] https://tinyurl.com/5n6nexze

5.1 Models

For fair comparison between existing LVLMs, each publicly available model we evaluate uses an LLM with ∼7B parameters. The LVLMs generally have different sized image encoders–*e.g.* LLaVA [30] uses a CLIP ViT-L/14 [9, 38] with an input resolution of 336 × 336, while InstructBLIP [7] uses a ViT-g/14 [50] trained with EVA [13] and an input resolution of 224 × 224. Note that image encoder size and resolution is not something we can easily control in a pre-trained model. Each model we consider contains instruction tuning in the final training phase—instruction tuned models provide free-form descriptions; THRONE focuses on models that *can* generate free-form descriptions. *E.g.* we leave out BLIP-2 [25] (response median length 31 characters) in favor of InstructBLIP (median response length 525).

5.2 THRONE Results on COCO

Results are shown in Tab. 1. The principal metric that we use to judge model performance is the classwise $F^{0.5}$-score (highlighted gray). We also report all the metrics outlined in Sec. 3.1, (P, $R, F^1, F^{0.5}$) for overall (left) and classwise averaging (right), utilizing the *unanimous* voting presented in Sec. 3.2. See the Supplementary Material for results and an analysis of different voting mechanisms. These results demonstrate that improvements on other benchmarks (POPE, MME, MMBench etc.) may be orthogonal and potentially at odds with improved performance on THRONE. Using the results of the 11 LVLMs that we evaluate, THRONE and POPE, which measure Type I and Type II hallucinations, respectively, have a Spearman's rank correlation coefficient of just 0.2, and THRONE vs POPE-C has just 0.4—the relationship between performance on POPE and THRONE *on the same dataset* is far from monotonic. For class-wise precision—P_{CLS}, the best performing models hallucinate ∼20% of the objects. We show in the Supplementary Material that the vast majority of false positive objects in the free-form image descriptions evaluated are direct hallucinations rather than misclassifications of visually similar objects (*e.g.* mistaking a squash racket for a tennis racket). These results demonstrate that much work still remains to adequately suppress Type I hallucinations in LVLMs.

5.3 THRONE Results on Objects365

Many LVLMs train on COCO directly or indirectly, thus to demonstrate generality we apply THRONE to the Objects365 dataset [42]. Like COCO, Objects365 aims to be exhaustive in its image-level class labeling (it aims to label at least one instance for each class present), but it has a larger object vocabulary and is not used as training data for the LVLMs that we evaluate. To gather a manageable subset of the Objects365 validation set (80k images), we use the natural sampling algorithm from [21], resulting in 5110 im-

Model	P_{ALL}	R_{ALL}	F^1_{ALL}	$F^{0.5}_{\text{ALL}}$	P_{CLS}	R_{CLS}	F^1_{CLS}	$F^{0.5}_{\text{CLS}}$
Adapter-v2 [15]	46.7	33.9	39.3	43.4	48.9	28.5	36.0	42.8
Adapter-v2.1 [15]	46.8	34.0	39.4	43.5	48.8	28.8	36.2	42.8
InstructBLIP [7]	54.5	37.2	44.2	49.8	53.7	33.6	41.3	48.0
Otter-Image [23]	21.4	12.7	16.0	18.8	9.5	4.4	6.0	7.7
MiniGPT4 [54]	53.0	32.9	40.6	47.2	49.9	31.9	39.0	44.9
MiniGPT-v2 [6]	54.5	36.0	43.4	49.4	51.3	*34.6*	41.3	46.8
mPLUG-Owl [48]	43.7	33.4	37.8	41.2	48.2	29.0	36.2	42.6
LRV-Instruction-v2 [28]	57.5	26.7	36.5	46.7	51.4	26.6	35.1	43.3
LLaVA-v1.3* [30]	*57.6*	32.9	41.9	50.1	52.6	30.5	38.6	45.9
LLaVA-v1.5 [29]	54.0	**39.5**	*45.6*	*50.3*	*53.9*	34.3	*41.9*	*48.4*
LLaVA-Mistral [19, 31]	**58.3**	*39.1*	**46.9**	**53.1**	**57.8**	**35.9**	**44.3**	**51.5**

Table 2. **THRONE Evaluation with Objects365.** Evaluation results for a selection of instruction-tuned LVLMs, we use a subset of Objects365 for the THRONE evaluation. Best and second-best performance are denoted by **blue** and *red*, respectively. *Our implementation using official code to enable fair comparison.

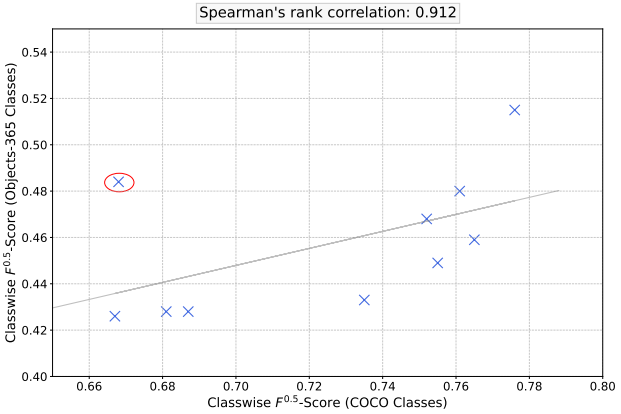

Figure 6. **Comparing THRONE on COCO and Objects365.** We observe that despite variations in the data distribution, THRONE metrics, which measure Type I hallucinations generalize and have a strong Spearman's rank correlation of $r = 0.900$. One model (red circle) designated as an outlier and ignored when calculating ranking correlation.

ages (the COCO validation set has 5k images). We present the results for THRONE on Objects365 in Tab. 2. Figure 6 shows the strong correlation in THRONE performance between evaluating on COCO and Objects365. This demonstrates that measuring Type I hallucinations in an LVLM using THRONE with a relatively small dataset like COCO, is indeed indicative of the intrinsic level of Type I hallucination in a given LVLM.

5.4 Completing POPE for Type II Hallucinations

Our experiments have used THRONE to evaluate the prevalence of Type I hallucinations in LVLM responses on COCO. POPE evaluates Type II hallucinations on COCO, but we find POPE is largely incomplete. First, POPE only evaluates on 500 COCO validation set images. Second, for each image only a subset of classes (at most 12) are evaluated—each image is only queried with 15% of possible questions. Finally, POPE artificially balances evaluation questions between positives and negatives, despite object class existence in images being inherently imbalanced.

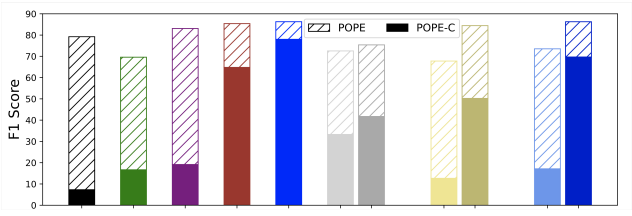

Figure 7. **Instability of POPE to complete evaluation of Type II Hallucinations.** Extending POPE to an exhaustive analysis on all COCO images and classes (POPE-C) leads to a dramatic reduction in performance across all 11 models. POPE is sensitive to the sampling mechanism, and by undersampling negative classes, POPE severely underestimates Type II hallucinations in LVLMs.

The above reasons make the evaluation of Type II hallucinations using POPE insufficient (see Sec. 2 for more details on POPE). To correct this, we complete POPE by using all images to exhaustively query each LVLM *for every class* in the COCO vocabulary, as done in THRONE. We name this POPE-Complete (POPE-C). Fig. 7 shows the extreme difference in evaluation between POPE and our exhaustive version, POPE-C—reporting F^1-score (POPE does not utilize $F^{0.5}$-score). As POPE only evaluates at most 9 negative classes, only a small subset of potential hallucinations of COCO classes are evaluated, thereby heavily underestimating the extent of Type II hallucination. For each LVLM analyzed, we observe a large—in many cases an extreme—reduction in precision and therefore in F1-score.

Our evaluation contains three pairs of LVLMs in which one is the follow-up work to the other, where each follow-up work generally trains on more data for more tasks with a more advanced language model. Comparing the right hand side of Fig. 7 to Tab. 1, we observe that these follow-up works generally show an improvement in POPE (and POPE-C), but surprisingly indicate a small reduction in performance on THRONE with COCO. This observation suggests that progress in reducing Type II hallucinations can be orthogonal to reducing Type I hallucinations.

5.5 Ablations

In the Supplementary Material we present three key ablation experiments and give an executive summary of results here.

First, after subsampling COCO images and LVLM responses, we replace the LMs in THRONE with human judgement as an oracle for Type I hallucination occurrence. When comparing THRONE and CHAIR with human judgements, we estimate using THRONE improves the precision of judging Type I hallucinations to 96% versus 91% when using CHAIR—this *reduces the false discovery rate by more than 50%*. Note that we find most estimated errors in THRONE arise from the particular class definitions in COCO, *e.g.* the COCO class `tv` includes computer monitors, which the oracle judgement is aware of.

Second, we apply the same class (and image) sampling

strategy as in POPE to THRONE and show this sampling overestimates $F^{0.5}_{\text{CLS}}$ by an average of 12.3 points compared to the complete use of classes in THRONE (Tab. 1).

Finally, we vary the choice of k *i.e.* the voting mechanism used to combine answers from multiple (LM, question) pairs. We use the *unanimous* voting mechanism ($k = 9$) in THRONE to minimize the false discovery rate and find the valid alternatives of *simple majority* ($k = 5$) or *all-but-one* ($k = 8$) voting mechanisms have strong correlations and rank correlations across all compute metrics, in THRONE, of > 0.99 and > 0.94, respectively.

6 Improved Baselines

Much needs to be done to study Type I hallucinations. As a first step to their mitigation, we demonstrate a baseline method to augment the visual instruction tuning data for LLaVA models [29, 30], yielding improvements on THRONE while maintaining similar performance regarding Type II hallucinations on POPE.

6.1 Visual Instruction Tuning Data Augmentation

Similar to chain-of-thought learning [46], during instruction tuning, we augment all visual instruction tuning samples constructed by LLaVA [30] by prepending the task of enumerating a list of objects (present and absent) and indicating approximate locations, if applicable. Other than this, the LLaVA data and training pipeline remains unchanged. To generate the new data for this object enumeration task, we use the same COCO bounding box annotations used to generate the vision instruction tuning data in LLaVA. The simple text-only format (no special tokens) we use is:

```
Instruction: <image> Give a list of objects and locations
             in the image.
Response:    {class_name_1} [{location_1}/absent]
             ...
             {class_name_N} [{location_N}/absent]
```

where `location_i` is a plain text indicator representing the location of the center point of the relevant object in the image on a 3×3 grid (*e.g.* `bottom left`). To provide negatives in the training data, if `class_name_i` is not present, we use the plain text indicator `absent`. Prior work [53] shows that classes that frequently co-occur in the training data are the most common hallucinations, therefore we bias our negative sampling towards class pairs that frequently co-occur using a correlation matrix. We detail and ablate this choice in the Supplementary Material.

6.2 Improved Baseline Results

Tab. 3 shows the result of evaluating our improved baseline method on THRONE, POPE, and POPE-C. During inference, we approximate our training data augmentation by first prompting the LVLM to perform the object enumeration tasks *and then* generating a response to the prompt from the relevant benchmark. We additionally show results of utilizing VisualGenome bounding box annotations [20] on the COCO images where available. On THRONE, we observe a large increase in classwise precision and therefore $F^{0.5}_{\text{CLS}}$, particularly for LLaVA-v1.5, demonstrating the ability of our method to reduce Type I hallucinations. Moreover, on POPE and POPE-C, using our object enumeration yields small improvements in precision, indicating reduced Type II hallucinations as well. In the Supplementary Material, we ablate the sampling of negatives during object enumeration training and the effect of removing the object enumeration task during inference.

Model	Object Enumeration Data	THRONE P_{CLS}	THRONE R_{CLS}	THRONE $F^{0.5}_{\text{CLS}}$	POPE P	POPE R	POPE F^1	POPE-C P	POPE-C R	POPE-C F^1
LLaVA-v1.3	✗	79.9	65.3	76.5	58.0	**98.4**	73.0	7.7	**99.2**	14.3
	COCO	83.2	**68.8**	79.9	73.2	88.2	80.0	9.8	69.4	17.2
	COCO + VG	**86.2**	67.0	**81.5**	**83.0**	82.5	**82.8**	**13.8**	50.4	**21.7**
LLaVA-v1.5	✗	69.9	56.4	66.8	81.9	90.8	86.1	58.7	85.7	69.7
	COCO	**87.2**	76.6	**84.9**	88.6	**85.3**	**87.0**	58.9	**87.5**	70.4
	COCO + VG	86.1	**77.0**	84.1	**89.8**	83.7	86.7	**64.5**	86.1	**73.7**

Table 3. **Improved Baseline via Object Enumeration:** Adding our object enumeration task to LLaVA training and inference leads to large improvements on THRONE particularly in terms of classwise precision, P_{CLS}, over standard LLaVA models, demonstrating a reduction in Type I hallucinations, as well as small reductions in Type II hallucinations judged by POPE and POPE-C.

7 Conclusion

We establish a novel benchmark, THRONE, for evaluating hallucinations generated by LVLMs in free-form image descriptions *i.e. Type I hallucinations*. Our benchmark utilizes multiple LMs and prompt formats with a simple voting mechanism to yield an accurate evaluation of Type I hallucinations in LVLM responses. We ensure that THRONE is broadly accessible by utilizing open-source LMs capable of running on a single commercial GPU. Using THRONE, we benchmark 11 publicly available LVLMs on two datasets, COCO and Objects365, and demonstrate that limited progress has been made in addressing Type I hallucinations. Moreover, we show how the established benchmark, POPE, underestimates *Type II hallucinations*, which occur in response to specific questions *e.g.* yes-no questions. We present results for a completed version (POPE-C) to enable a comparison of Type I hallucinations through THRONE and Type II hallucinations using POPE-C. Finally, we propose a simple data augmentation for LVLM training that can result in a large reduction in Type I hallucinations whilst maintaining or improving Type II hallucination performance.

Limitations and Ethical Considerations are discussed in the Supplementary Material.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in*

Neural Information Processing Systems, 35:23716–23736, 2022. 1, 3

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the International Conference on Computer Vision*, pages 2425–2433, 2015. 3

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 3

[4] Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. Visit-bench: A benchmark for vision-language instruction following inspired by real-world use, 2023. 2

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 3

[6] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 1, 6, 7

[7] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. 1, 3, 6, 7

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 5

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *Proceedings of the International Conference on Learning Representations*, 2021. 6

[10] Mark Everingham, Luc Van Gool, Chris K. I. Williams, John Winn, and Andrew Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *International Journal of Computer Vision*, 2010. 4

[11] Hugging Face. Text generation inference. https://github.com/huggingface/text-generation-inference, 2023. Accessed: November 10, 2023. 6

[12] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015. 3

[13] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 6

[14] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 1, 3

[15] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 6, 7

[16] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations*, 2022. 3

[17] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023. 1

[18] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning*, pages 4904–4916, 2021. 3

[19] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 6, 7

[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2017. 8

[21] Kibok Lee, Hao Yang, Satyaki Chakraborty, Zhaowei Cai, Gurumurthy Swaminathan, Avinash Ravichandran, and Onkar Dabeer. Rethinking few-shot object detection on a multi-domain benchmark. In *Proceedings of the European Conference on Computer Vision*, 2022. 6

[22] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 1, 2, 3

[23] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023. 6, 7

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 1

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, 2023. 1, 3, 6

[26] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language*, 2023. 1, 2, 3, 4

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 2014. 2, 3, 4, 5

[28] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 6, 7

[29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 3, 6, 7, 8

[30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. 1, 3, 6, 7, 8

[31] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 6, 7

[32] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 1, 2, 3, 5

[33] Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning. In *Proceedings of the International Conference on Machine Learning*, 2023. 5

[34] Hwee Tou Ng, Siew Mei Wu, Ted Briscoe, Christian Hadiwinoto, Raymond Hendy Susanto, and Christopher Bryant. The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*, 2014. 5

[35] OpenAI. Gpt-4 technical report, 2023. 2, 3

[36] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 1, 3

[37] Tina D Purnat, Paolo Vacca, Christine Czerniak, Sarah Ball, Stefano Burzo, Tim Zecchin, Amy Wright, Supriya Bezbaruah, Faizza Tanggol, Ève Dubé, Fabienne Labbé, Maude Dionne, Jaya Lamichhane, Avichal Mahajan, Sylvie Briand, and Tim Nguyen. Infodemic signal detection during the covid-19 pandemic: Development of a methodology for identifying potential information voids in online conversations. *JMIR Infodemiology*, 1(1):e30971, 2021. 5

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 6

[39] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020. 3, 5

[40] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language*, pages 4035–4045, 2018. 1, 2, 3

[41] Giuliano Rossi, Jakub Kolodziej, and Gurvinder Brar. A recommender system for active stock selection. *Computational Management Science*, 17, 2020. 5

[42] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the International Conference on Computer Vision*, 2019. 5, 6

[43] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 3

[44] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, S. M. Ali Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. In *Advances in Neural Information Processing Systems*, pages 200–212, 2021. 1, 3

[45] 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. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023. 1

[46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 2022. 8

[47] Hongbin Ye, Tong Liu, Aijia Zhang, Wei Hua, and Weiqiang Jia. Cognitive mirage: A review of hallucinations in large language models. *arXiv preprint arXiv:2309.06794*, 2023. 1, 3

[48] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 1, 3, 6, 7

[49] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 3

[50] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 6

[51] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. Siren's song in the ai ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023. 1, 3

[52] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. 1

[53] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models, 2023. 1, 8

[54] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 6, 7