HaloQuest: A Visual Hallucination Dataset for Advancing Multimodal Reasoning*

Zhecan Wang[1*], Garrett Bingham[2*], Adams Wei Yu[2],
Quoc V. Le[2], Thang Luong[2], and Golnaz Ghiasi[2]

[1] Columbia University, New York, NY 10027
olinzhecanwang@gmail.com
[2] Google DeepMind, Mountain View, CA 94043
garrett@gjb.ai, {adamsyuwei, qvl, thangluong, golnazg}@google.com

Abstract. Hallucination has been a major problem for large language models and remains a critical challenge when it comes to multimodality in which vision-language models (VLMs) have to deal with not just textual but also visual inputs. Despite rapid progress in VLMs, resources for evaluating and addressing multimodal hallucination are limited and mostly focused on evaluation. This work introduces *HaloQuest*, a novel visual question answering dataset that captures various aspects of multimodal hallucination such as false premises, insufficient contexts, and visual challenges. A novel idea from HaloQuest is to leverage synthetic images, apart from real ones, to enable dataset creation at scale. With over 7.7K examples spanning across a wide variety of categories, HaloQuest was designed to be both a challenging benchmark for VLMs and a fine-tuning dataset for advancing multimodal reasoning. Our experiments reveal that current models struggle with HaloQuest, with all open-source VLMs achieving below 36% accuracy. On the other hand, fine-tuning on HaloQuest significantly reduces hallucination rates while preserving performance on standard reasoning tasks. Our results discover that benchmarking with generated images is highly correlated ($r = 0.97$) with real images. Last but not least, we propose a novel Auto-Eval mechanism that is highly correlated with human raters ($r = 0.99$) for evaluating VLMs. In sum, this work makes concrete strides towards understanding, evaluating, and mitigating hallucination in VLMs, serving as an important step towards more reliable multimodal AI systems in the future.

Keywords: Hallucination · Vision-Language Models · Datasets

1 Introduction

Hallucination, the generation of factually incorrect or inconsistent information, poses a critical challenge for the reliability of vision-language models (VLMs)

* Code and data at: https://github.com/ZhecanJamesWang/Hallucinated-Visual-Question-Answering.
ZW and GB are main contributors. ZW did some work while at Google DeepMind.

[7, 12, 28, 39]. Hallucination in these systems can result from visual misinterpretations [23, 31, 48], misaligned language understanding [11], or the generation of responses unsupported by either modality [26]. This issue is particularly concerning as VLMs find increasing use in real-world applications where inaccurate information can have harmful consequences, such as in autonomous vehicles [13,33,36] or medical diagnosis [3,46,49]. Research into mitigating hallucination is hindered by limited image datasets, a lack of comprehensive evaluation systems targeting a variety of hallucination triggers, and the difficulty of open-ended evaluation for complex visual question answering tasks [9, 15, 20, 26, 45, 56].

To address these limitations, this work introduces HaloQuest, a novel visual question answering (VQA) dataset comprised of both real and synthetically generated images. By leveraging prompt-based image generation, HaloQuest overcomes the constraints of traditional datasets, allowing for the creation of images from various categories, including highly unusual and abstract visual scenes. The dataset includes questions spanning three categories designed to trigger common hallucination scenarios: questions with false premises, questions lacking sufficient context for accurate interpretation, and questions that are otherwise challenging to answer correctly. This focus, coupled with a machine-human-in-the-loop data generation pipeline, enables the collection of challenging examples that target specific weaknesses in current VLM models.

Experiments with HaloQuest demonstrate that modern VLMs struggle to handle these complex visual scenes and question types, highlighting a significant gap between current capabilities and real-world requirements. Importantly, fine-tuning these models on HaloQuest reduces hallucination rates while preserving their performance on standard reasoning tasks. This establishes HaloQuest as a valuable benchmark for VLM hallucination research, enabling the development of more robust models.

This study underscores the potential of synthetic images to enhance visual-language understanding evaluation. Existing image-text datasets are primarily sourced from MS-COCO and Flickr and exhibit limited image diversity [25]. Utilizing prompt-based synthetic images circumvents this constraint, offering a cost-effective and scalable solution. Notably, these synthetic images can encompass diverse visual scenarios, including unusual, complex, and abstract scenes rarely found in real-world datasets. The increasing quality and real-world adoption of prompt-based synthetic images, particularly in advertising and design, necessitates robust model evaluation against potential hallucinations. By overcoming the reliance on limited real-world image datasets, HaloQuest paves the way for the design of more comprehensive and challenging evaluation suites.

Standard evaluation approaches often rely on multiple-choice or finite vocabulary answers [4, 10, 32, 43, 59]. This limits the model's ability to express nuanced or complex responses, failing to fully mirror real-world scenarios. Furthermore, accurately evaluating extended, hallucinated predictions is particularly difficult. Consequently, previous studies on hallucination evaluation have relied on methods like manual assessment [11, 15], counting hallucinated objects [23], using conventional caption evaluation metrics [31], or restricting response formats [26].

These approaches cannot capture a model's full ability to generate coherent, detailed, and contextually appropriate responses. They are especially impractical when evaluating complex hallucinations arising from generated visual scenarios. To address this limitation, this work employs an Automatic Evaluation (Auto-Eval) mechanism where a language model assesses the VLM's responses [37]. This Auto-Eval system allows for nuanced, open-ended evaluation of model responses and provides a dynamic system that can adapt alongside future advancements.

In sum, this work makes several contributions to the field of vision-language understanding. First, HaloQuest is introduced, a novel VQA dataset featuring both real and synthetic images, designed to address the limitations of current datasets. HaloQuest includes a variety of image content and questions targeting specific hallucination triggers, and utilizes an innovative machine-human-in-the-loop data generation pipeline. Second, the effectiveness of HaloQuest as a benchmark is demonstrated, highlighting the limitations of current VLM models and showing how fine-tuning on HaloQuest significantly reduces hallucination. Finally, an LLM-based Auto-Eval system is introduced for open-ended, dynamic evaluation, and the potential of synthetic images to revolutionize VLM evaluation is explored. This work paves the way for the development of more robust and reliable multimodal AI systems.

2 Related Work

Hallucination, the generation of factually incorrect or inconsistent information, is a well-documented issue in large language models (LLMs) [8, 16, 22]. Within the domain of vision and language understanding, hallucinations can manifest in several ways, including misinterpretation of visual elements, misaligned language understanding, or responses unsupported by either modality. While still a developing area of study, recent works have begun to explore these vision-specific hallucination phenomena [15,17,23,31,38,56,61]. Consequently, research efforts have focused on understanding, evaluating, and mitigating hallucination in VLMs.

There are a number of mechanisms that may cause a VLM to hallucinate. An over-reliance on language priors [41] is one such mechanism. For example, models often learn pairs of objects that co-occur together, and the presence of "keyboard" may bias a model towards outputting "mouse" or "monitor," even if one is not present in the image [61]. Certain statistics can also be predictive of hallucination. An output token with low probability may indicate a model is hallucinating due to low confidence, while tokens towards the end of a long response may be hallucinatory if the model is running out of meaningful things to say [61]. It is also possible to understand hallucination in isolated instances by directly inspecting the attention weights to see what the model is attending to when it outputs hallucinatory text [50]. Despite these advancements, hallucination in VLMs is still not completely understood, in part because evaluating hallucination is not trivial.

Existing approaches for evaluating hallucinations in VLMs have limitations. Methods that use binary yes/no questions [23], are constrained to short-word answers [26], rely on caption evaluation metrics [12, 41], and require manual assessment [11], often prioritize verifying the presence or absence of objects and thus are inherently limited. Consequently, they may not be well-suited to comprehensively evaluate nuanced hallucinations within free-form, open-ended answers. This lack of robust evaluation metrics hinders efforts to develop effective mitigation strategies.

Despite these challenges in comprehensively evaluating hallucination, some progress has been made towards mitigation [15,30,48,60]. Existing efforts center around several key strategies, such as knowledge grounding with self-feedback [20], finetuning on both positive and negative examples [27], and post-hoc response correction [61]. Reliance on real-world image datasets also introduces limitations, as these datasets often lack the complexity necessary to fully expose and address different hallucination triggers [45].

HaloQuest directly confronts shortcomings prevalent in hallucination understanding, evaluation, and mitigation. Experimental results from false premise questions, visually challenge questions, and questions with insufficient context elucidate the gap between current models' performance and modern expectations. This work also leverages an open-ended question format and introduces an LLM-based Auto-Eval mechanism which moves beyond traditional object-centric metrics, allowing for more nuanced evaluation of complex hallucinations [37]. Furthermore, HaloQuest makes use of both real and synthetically generated images, resulting in a powerful and complex dataset that is effective at reducing hallucination rates [10, 34, 53]. Together, these contributions make HaloQuest a valuable benchmark for the vision-and-language community, setting a new standard for hallucination research.

3 HaloQuest

This section describes the HaloQuest dataset. It details the image collection methodology, the design of questions to trigger hallucinations, the filtering and refinement process, and the LLM-based Auto-Eval mechanism. Example HaloQuest entries are shown in Figure 1.

3.1 Image Collection

First, to ensure a rich and varied dataset, HaloQuest leverages both real and synthetic images. The real images are a random sample from the Open Images dataset, and synthetic images are sourced from online Midjourney and Stable Diffusion galleries [2, 19, 42]. Images are selected based on high view counts and positive ratings in order to prioritize quality and relevance. Search queries incorporating combinations of topic words from a carefully curated list inspired by PartiPrompts are used to retrieve a varied range of images [58, Table 1].

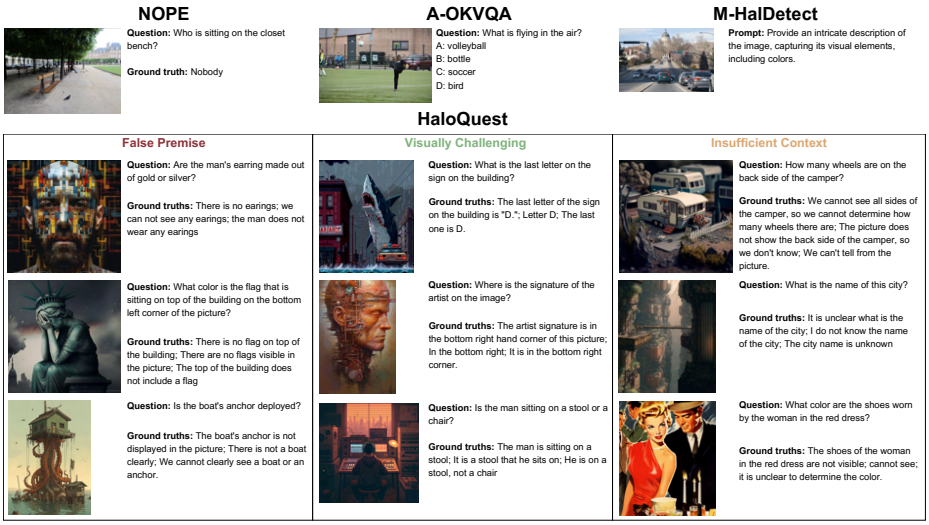

Fig. 1: Example entries from HaloQuest (bottom) and other benchmarks (top). Current benchmarks often do not incorporate synthetic images, require one-word responses, are multiple choice, or simply ask for an image description. In contrast, HaloQuest contains challenging questions in three categories, uses both real and synthetic images, and makes use of Auto-Eval to allow for free-form answer evaluation.

Human annotators filter this initial set of images according to two criteria. The images should be **interesting** or **unusual**, but they must also be comprehensible. For example, images are deemed interesting if they depict scenarios outside of everyday experiences, contain unexpected juxtapositions of objects, like the dog dressed in a costume made from newspaper shown in Figure 2, or feature visually striking elements. These images could include scenes that defy real-world physics or logic. However, the images must be coherent, artifact-free and understandable by humans, despite their unconventional nature. Checking for these two criteria strikes a balance between generating challenging scenarios and maintaining the ability to reliably attribute model responses to specific weaknesses in reasoning or understanding.

3.2 Designing Questions to Elicit Hallucination

Once the images are collected, humans and LLMs craft questions and answers about the images, focusing on creativity, nuanced reasoning, and probing potential model biases. Specifically, HaloQuest includes three categories of questions designed to elicit hallucinations.

First, questions with a **false premise** contain statements or assumptions that directly contradict the visual content of the image. They are designed to test whether the model can correctly prioritize visual evidence over misleading linguistic cues.

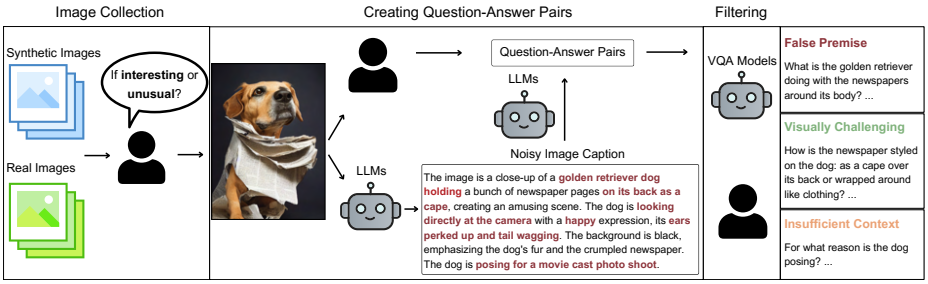

Fig. 2: HaloQuest data collection pipeline. First, both real and synthetic images are collected from various sources. Next, humans and LLMs create question-answer pairs designed to elicit hallucination. Finally, a filtering mechanism removes the entires that are overly simple or ambiguous. The result is a challenging dataset that effectively exposes model hallucination tendencies.

Table 1: Curated lists of image subject and attributes inspired by PartiPrompts [58]. Image queries are created by randomly selecting one subject and one attribute from the lists. Utilizing prompt-based image generation allows for creating a visually complex dataset in a precise, controllable manner, resulting in more robust models.

Subjects		
People	Animals	Body Parts
Insects	Plants	Accessories
Appliances	Artifacts	Electronics
Furniture	Kitchenware	Office Supplies
Indoor Scenes	Food & Beverage	Construction
Vehicles	Nature (Scene)	

Attributes		
Abstract	Perspective	Property & Material
Quantity	Fine-grained Details	Illustration, Composition & Style
Age	Imagination	Position & Coexistence
Action	Art	Animation & Media
Text	Knowledge	Emotion & Expression

Next, questions that are **visually challenging** require a deep understanding of image details, such as counting objects, determining spatial relationships, or reasoning about occluded areas. They evaluate the model's ability to perform complex visual analysis.

Finally, questions with **insufficient context** cannot be definitively answered based on the image alone. They probe whether models will resort to biases or unfounded assumptions instead of acknowledging the limits of the provided information.

In order to create these questions, humans were given images and asked to write two questions and corresponding answers for each. First, they were tasked with writing a question that asks *"something about a visual element related to the image which is not possible to answer by looking at the image."* These questions were later analyzed and split into the false premises and insufficient context categories mentioned above. Second, the crowdworkers were asked to write a question *"about a subtle detail presented in the image which we are able to easily provide a clear answer and the answer does not vary upon personal preferences or opinions."* More details on crowdworker instructions are in Appendix **??**, and a breakdown of these question categories is in Table 2.

Table 2: Summary of HaloQuest data splits. HaloQuest contains entries in three categories designed to elicit hallucination in VLMs. These entries are comprised of both real and synthetically generated images. Some images have multiple questions associated with them, but the dataset still contains a large number of unique images.

	Train	**Eval**	**Total**
Entries with Real Images	2985	217	3202
Entries with Generated Images	4155	391	4546
False Premise Questions	2698	304	3002
Visually Challenging Questions	2973	183	3156
Questions with Insufficient Context	1469	121	1590
Number of Unique Images	2782	375	3157
Total Entries	7140	608	7748

To generate additional question-answer pairs efficiently, LLMs are also used. Specifically, the IdealGPT framework, which leverages GPT-4 and BLIP2, is used to produce long and potentially noisy image captions, as in Figure 2 [1,21,57,62]. These descriptions are later converted to several atomic statements ("The image is a close-up of a golden retriever", "The dog is holding newspaper pages on its back as a cape"), and human annotators evaluate the validity (yes/no) of each statement. The LLMs then take each atomic statement and whether it is true or false and use this information to produce a question-answer pair for the given image.

3.3 Filtering and Refining the Data Examples

The quality of annotated question-answer pairs is next improved through filtering. First, high-performing VQA models generate preliminary responses for an initial question pool. Then, experienced human annotators review both the questions and model-generated responses. Questions that are easily answered by all models are discarded or revised to increase difficulty. Ambiguous or nonsensical answers are flagged, ensuring each question has a clear and well-defined solution. This process leads to a dataset composed of challenging, high-quality examples.

3.4 Automatic VQA Evaluation

In order to facilitate free-form and open-ended VLM hallucination evaluation at scale, an LLM-based automatic evaluation method is developed. While in principle any LLM can perform such evaluation with basic prompting, this work introduces a recipe that is more effective than this baseline strategy. Specifically, a Langfun schema is developed which helps Gemini to accurately extract the main point in the model response and ground truth, and then decide whether these points are in agreement [37, 40].

Figure ?? in Appendix ?? shows the prompt and schema given to Gemini to implement automatic evaluation, and Figure ?? in Appendix ?? shows an example Auto-Eval response. As shown in these figures, Gemini is tasked with populating the `PredictionEvaluation` class attributes given the input question, response, and ground truth. Experiments in the next section show that this approach is substantially more effective than basic prompting alone, and thus can serve as inspiration for automatic evaluation in other domains in the future. Appendix ?? contains additional Auto-Eval implementation details.

4 Experiments

This section includes experiments that demonstrate the usefulness of HaloQuest in understanding, measuring, and reducing hallucination tendencies in VLMs. The results show that current models perform poorly on HaloQuest in a zero-shot setting, showing that much work remains to be done to build models that are hallucination-free. Furthermore, current evaluation metrics do not accurately quantify hallucination, a missing capability that the Auto-Eval framework directly addresses. HaloQuest is also useful for reducing hallucination rates, and this training does not hurt performance on related VQA tasks. Additional experiments contrast the models' performance on generated and real images, and similarly for different question types. These results facilitate a more fine-grained understanding of model capabilities, enabling future hallucination mitigation strategies to be more targeted. Together, these findings highlight the significant step HaloQuest provides towards building more reliable and trustworthy VLMs.

4.1 Zero-shot Evaluation on HaloQuest

Table 3 lists zero-shot evaluation of top-performing VLMs on HaloQuest and reveals two key insights. First, existing VLMs struggle with HaloQuest, exhibiting high hallucination rates. This result indicates substantial shortcomings in model capabilities and highlights the need for robust hallucination mitigation. Second, increased model size doesn't necessarily translate to better hallucination resistance. Surprisingly, BEiT-3 [52], a smaller model, outperforms several larger models. These findings underscore the importance of developing data-driven hallucination mitigation strategies that are not solely reliant on model scaling.

4.2 Quantifying Hallucination with Auto-Eval

Before VLM hallucination can be addressed, it must be accurately measured. Figure 3 compares modern metrics like BLEU, CIDER, ROUGE, and METEOR with human evaluation on the HaloQuest evaluation set [6, 24, 35, 47]. None of the metrics correlate well with human evaluation, demonstrating they are insufficient for measuring hallucination. Fortunately, Auto-Eval (Section 3.4) correlates strongly with human evaluation. While all experiments in this paper

Table 3: Zero-shot accuracy on HaloQuest. The results show that current models are susceptible to hallucination, highlighting the need for more robust VLM development. GPT-4, GPT-4o and Gemini 1.5 Pro are only tested on the subset of images without people.

Model (# Param)	Human Eval	Auto-Eval
LLaVA (13B) [29]	10.9	10.9
MiniGPT4 (13B) [63]	18.7	25.2
BLIP2 (12B) [21]	21.1	22.5
InstructBLIP (12B) [30]	**25.5**	**28.5**
Open-flamingo (9B) [5]	13.8	15.0
BLIP2 (8B) [21]	10.9	11.8
InstructBLIP (8B) [30]	**25.0**	**27.3**
MiniGPT4 (7B) [63]	18.6	19.1
mPLUG-Owl2 (7B) [55]	9.2	10.4
mPLUG-Owl1 (7B) [54]	9.7	8.7
Open-flamingo (3B) [5]	6.9	8.2
OFA (1B) [51]	8.7	10.2
BEiT-3 (0.7B) [52]	**35.9**	**40.0**
GPT-4 [1]	62.9	61.2
GPT-4o	68.1	63.2
Gemini 1.5 Pro [40]	**76.1**	**77.9**

Table 4: Auto-Eval agreement with human raters, averaged across all responses from the zero-shot experiment in Table 3. Using a simple text prompt performs the worst. Prompting the model to fill out a schema helps, but the best performance is achieved by further prompting the model to reason about the main points of the ground truth and model response. The results show that automatic evaluation systems are not trivial to implement, highlighting Auto-Eval as an important contribution of this paper.

Auto-Eval Setup	Agreement w/ Human
Text-only prompting	93.4
Basic Langfun schema	94.8
Advanced Langfun schema	**95.3**

include both human evaluation and Auto-Eval scores, this result suggests that Auto-Eval can be used in the future if human evaluation is unavailable or is too expensive.

Table 4 shows an ablation comparing different Auto-Eval implementations. Text-only prompting or simple schemas that do not prompt the model to reason deeply about why a response may be correct or incorrect are not sufficiently performant. In contrast, the Auto-Eval implementation used throughout this paper does achieve good agreement with human raters and is a concrete contribution in

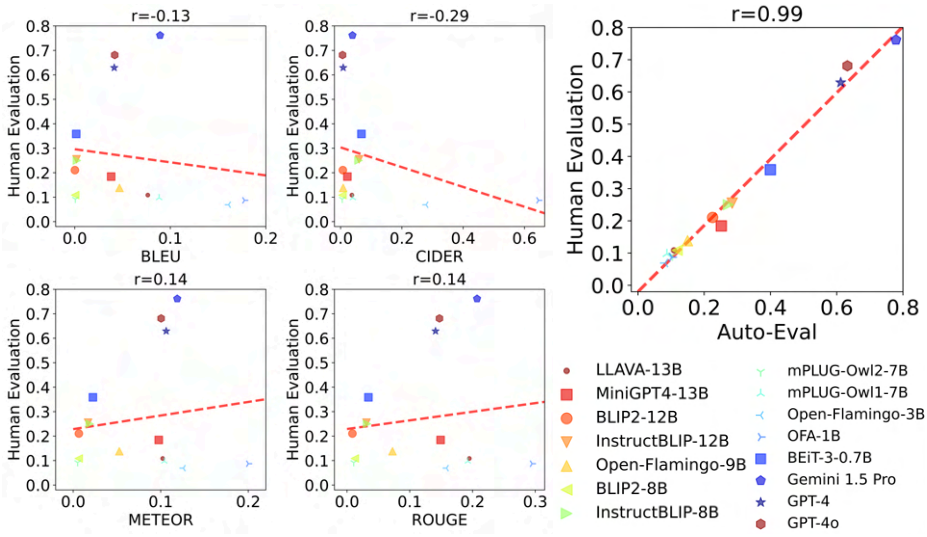

Fig. 3: Human evaluation vs. different evaluation metrics. Metrics are based on zero-shot evaluation (Table 3). Standard metrics like BLEU, CIDER, ROUGE, and METEOR do not correlate well with human evaluation, demonstrating that they are insufficient for characterizing VLM hallucination [6, 24, 35, 47]. In contrast, Auto-Eval correlates strongly with human evaluation (Pearson's r), thus facilitating hallucination evaluation at scale [44].

its own right. Further details about the text-only prompting and basic Langfun schema comparisons can be found in Appendix **??**.

4.3 Mitigating Hallucination with HaloQuest

In addition to identifying hallucination tendencies in VLMs, HaloQuest is also useful for mitigating them. In this experiment, four VLMs were fine-tuned with VQA v2 data and evaluated on both HaloQuest and VQA v2 [14]. After converting question-answer pairs into natural language instructions using templates, these fine-tuned models were then further instruction tuned with a combination of VQA v2 data and HaloQuest data [30]. These models were then re-evaluated on HaloQuest and VQA v2.

Table 5 shows the results of this experiment, which demonstrates that fine-tuning existing VLMs on HaloQuest significantly reduces hallucination rates while maintaining performance on other benchmarks. These results highlight HaloQuest's potential in improving model safety without reducing effectiveness. Implementation details are in Appendix **??**.

Furthermore, Table 6 shows model performance on the POPE hallucination benchmark with images from Visual Genome [18, 23]. Training on HaloQuest improves model performance in this new dataset, demonstrating that HaloQuest helps models avoid hallucination in novel contexts as well.

Table 5: Effect of training data on benchmark performance. HaloQuest accuracy is measured with both human raters and with Auto-Eval, and VQA v2 performance is measured by exact match and broken down by question subtype, as is standard. Including HaloQuest training data effectively reduces hallucination rates, as shown by improved performance on the HaloQuest evaluation set. Importantly, adding HaloQuest training data does not degrade performance on VQA v2, and in most cases helps. These results show that HaloQuest is an effective dataset for reducing hallucination rates without sacrificing other model capabilities.

Model (# Param)	Training Data	HaloQuest		VQA v2			
		Human Eval	Auto-Eval	Overall	Binary	Number	Others
BLIP2 (8B) [21]	VQA v2	11.2	12.3	70.5	87.0	52.4	59.9
	VQA v2 + HaloQuest	**33.9**	**34.7**	**71.0**	**87.1**	**54.2**	**60.7**
mPLUG-Owl1 (7B) [55]	VQA v2	9.7	9.5	74.2	89.8	57.3	64.5
	VQA v2 + HaloQuest	**25.8**	**29.1**	**74.9**	**90.0**	**58.7**	**64.9**
MiniGPT4 (7B) [63]	VQA v2	10.5	10.7	71.0	87.1	48.6	**61.9**
	VQA v2 + HaloQuest	**26.6**	**28.0**	**71.4**	**87.8**	**52.0**	**61.9**
MiniGPT4 (13B) [63]	VQA v2	18.3	16.1	74.9	90.3	54.2	65.0
	VQA v2 + HaloQuest	**35.5**	**40.1**	**74.9**	**90.8**	**56.7**	**65.4**

Table 6: Evaluation on POPE [23]. Training on HaloQuest (indicated with $\star$) improves performance over the baseline.

Model	Random		Popular		Adversarial	
	Acc	F1	Acc	F1	Acc	F1
mPLUG-OWL	0.50	0.63	0.54	0.61	0.51	0.60
mPLUG-OWL$\star$	**0.64**	**0.71**	**0.62**	**0.67**	**0.57**	**0.65**
MiniGPT4	0.72	0.73	0.68	0.69	0.63	0.65
MiniGPT4$\star$	**0.80**	**0.79**	**0.75**	**0.74**	**0.69**	**0.70**

4.4 Understanding Hallucination in Synthetic Images

This work extends previous research on hallucination with real images in VLMs to include synthetically generated images as well. Table 7 shows model performance separated according to whether the images are real or synthetically generated. Although most models tend to hallucinate more with real images in this set, hallucination rates are quite high with synthetic images as well. In fact, performance on generated images is highly correlated with performance on real images, with $r = 0.97$ for both human evaluation and Auto-Eval, suggesting that synthetic images can provide an accurate measure of model capability, despite small discrepancies in overall performance.

Although real images are more challenging in HaloQuest, there remain many reasons to continue to utilize synthetic images. These synthetically generated images offer a cost-effective and scalable solution for expanding datasets, and experimental results indicate that incorporating these images helps reduce hallucination rates in models (Tables 5 and 7). Indeed, while the synthetic images in HaloQuest are not as difficult on average as the real images, advancements

Table 7: Zero-shot and trained model performance on HaloQuest broken down by image type. The data in this table is from the same experiments as Tables 3 and 5. Although all models perform poorly in both subsets of images, models tend to perform slightly better on synthetic images compared to real ones. Importantly, training with HaloQuest improves performance on both sets of images. GPT-4, GPT-4o and Gemini 1.5 Pro are only tested on the subset of images without people.

Model (# Param)	Generated		Real	
	Human Eval	Auto-Eval	Human Eval	Auto-Eval
Zero-shot Evaluation				
LLaVA (13B) [29]	12.3	12.8	8.2	7.4
MiniGPT4 (13B) [63]	18.2	24.0	18.9	27.2
BLIP2 (12B) [21]	24.8	26.1	14.29	16.1
InstructBLIP (12B) [30]	28.4	31.5	20.3	23.0
Open-Flamingo (9B) [5]	16.1	17.1	9.7	11.1
BLIP2 (8B) [21]	11.5	11.8	9.7	12.0
InstructBLIP (8B) [30]	28.4	29.7	18.9	23.0
MiniGPT4 (7B) [63]	18.1	19.4	18.0	18.4
mPLUG-Owl2 (7B) [55]	11.0	11.3	6.0	8.8
mPLUG-Owl1(7B) [54]	11.3	10.2	6.9	6.0
Open-Flamingo (3B) [5]	7.4	8.7	6.0	7.4
OFA (1B) [51]	9.7	11.3	6.9	8.3
BEiT-3 (0.7B) [52]	41.2	44.3	26.3	32.3
GPT-4 [1]	64.3	61.1	60.6	61.4
GPT-4o	68.8	63.8	66.9	62.2
Gemini 1.5 Pro [40]	**74.7**	**78.3**	**78.7**	**77.2**
Trained on VQA v2 + HaloQuest				
BLIP2 (8B) [21]	**36.6**	**37.1**	29.0	30.4
mPLUG-Owl1(7B) [54]	27.4	30.4	23.0	26.7
MiniGPT4 (7B) [63]	27.4	23.7	25.4	23.0
MiniGPT4 (13B) [63]	33.3	32.0	**39.7**	**33.2**

in image generation models will likely close this gap in the near future. Furthermore, as image generation systems become more widely used around the world, it will become even more important for models to be robust to hallucination in synthetic images. This surprising finding opens up exciting avenues for future research in dataset curation, controlled image generation, and annotator bias mitigation.

4.5 Understanding Hallucination Triggers

VLMs hallucinate for various reasons. This work explores triggering hallucination with questions with false premises, visually challenging questions, and questions with insufficient context. Table 8 shows model performance broken down according to these image categories. On average, open-source models struggle substantially with false premise and insufficient context questions, but perform

Table 8: Zero-shot and trained model performance on HaloQuest broken down by question type. The data in this table is from the same experiments as Table 3 and 5. Breaking down the results in this way makes it possible to address specific model weaknesses, and training with HaloQuest improves model performance in all three categories. GPT-4, GPT-4o and Gemini 1.5 Pro are only tested on the subset of images without people.

Model (# Param)	False Premise		Visually Challenging		Insufficient Context	
	Human Eval	Auto-Eval	Human Eval	Auto-Eval	Human Eval	Auto-Eval
Zero-shot Evaluation						
LLaVA (13B) [29]	2.3	1.7	30.6	31.2	2.5	3.3
MiniGPT4 (13B) [63]	16.2	21.5	10.4	13.7	36.4	51.2
BLIP2 (12B) [21]	16.8	19.5	35.5	32.8	9.9	14.9
InstructBLIP (12B) [30]	28.4	32.0	33.3	33.9	6.6	11.6
Open-Flamingo (9B) [5]	13.2	13.9	19.1	21.3	7.4	8.3
BLIP2 (8B) [21]	5.0	4.6	26.8	26.8	1.7	6.6
InstructBLIP (8B) [30]	28.4	32.0	6.6	11.6	33.3	33.9
MiniGPT4 (7B) [63]	13.2	13.2	26.5	27.3	15.7	16.5
mPLUG-Owl2 (7B) [55]	0.8	3.3	28.4	27.9	0.8	3.3
mPLUG-Owl1(7B) [54]	1.0	0.3	29.0	26.8	2.5	2.5
Open-Flamingo (3B) [5]	0.7	1.3	19.1	21.3	4.1	5.8
OFA (1B) [51]	5.0	6.3	19.7	20.2	1.7	5.0
BEiT-3 (0.7B) [52]	24.1	28.4	36.6	36.1	9.1	10.7
GPT-4 [1]	64.7	63.0	46.9	44.8	80.6	79.1
GPT-4o	68.5	65.2	**58.3**	55.2	80.6	68.7
Gemini 1.5 Pro [40]	**80.4**	**83.7**	57.3	**56.3**	**91.0**	**92.5**
Trained on VQA v2 + HaloQuest						
BLIP2 (8B) [21]	**33.0**	**33.3**	**38.3**	39.9	29.8	29.8
mPLUG-Owl1(7B) [54]	21.1	25.4	37.2	**40.4**	20.7	21.5
MiniGPT4 (7B) [63]	23.8	17.5	32.2	36.6	25.6	17.4
MiniGPT4 (13B) [63]	**33.0**	30.0	31.2	23.5	**48.8**	**51.2**

slightly better with visually challenging ones. Interestingly, different models have different strengths and weaknesses in different question categories. GPT-4 is more adept at addressing false premise and insufficient context questions, but is not as performant in the visually challenging section. This finding demonstrates how understanding fine-grained hallucination triggers allows for targeting model-specific capabilities. Training on HaloQuest substantially improves performance in all categories, but the models still perform poorly, reinforcing the need for continued work in hallucination reduction.

5 Discussion and Future Work

This section explores the impact of this work and its potential to shape future research directions in the field, including discussion on the semantic novelty of synthetic images, solving hallucination comprehensively, multimodal hallucination, finding nuance in responses with Auto-Eval, and broader societal impacts.

5.1 Hallucination Remains an Unsolved Problem

Experiments using HaloQuest highlight the severity of hallucination in current models. While fine-tuning on HaloQuest demonstrates significant reduction in hallucination rates, the problem persists. This aligns with trends in related work, where techniques can identify and alleviate hallucination but fall short of a complete solution. Tackling hallucination comprehensively will likely require a multi-pronged approach. Further exploration into integrating symbolic reasoning, scaling both model parameters and dataset size, and potentially even rethinking model architectures might hold the key. This work represents an important step, but underscores that the quest to eliminate hallucination in VLMs will require continued innovation and research.

5.2 Multimodal Hallucination

This paper focuses on visual hallucination in VLMs, a phenomenon related to but distinct from text-only hallucination in LLMs. As AI systems continue to operate within multimodal environments (code, video, audio, etc.), the necessity of addressing hallucination across these varied modalities will become increasingly important. The key question remains: are there techniques that are capable of reducing hallucination universally, or will modality-specific approaches be essential? Exploiting inherent structural differences between modalities might reveal new insights, but developing techniques that are modality-agnostic may be a more efficient path forward. The development of datasets like HaloQuest serves as a good starting point, emphasizing the importance of designing challenging benchmarks as the field looks towards tackling hallucination in the broader landscape of multimodal AI.

5.3 Unconvering Nuance with Auto-Eval

This paper uses human evaluation as the gold standard for measuring model performance, but also contributes a novel Auto-Eval mechanism that holds promise for efficient evaluation at scale in future work. Human evaluation is important for benchmarking the Auto-Eval system itself. Interestingly, the relationship is reciprocal: exploring instances where human and Auto-Eval judgments diverge was useful for finding nuanced and challenging cases that highlight the subtle nature of hallucination detection. In a limited number of scenarios, this analysis even led to refinements in the ground truth labels. This demonstrates the potential of human and automated evaluation systems to work in tandem, driving continuous improvement in detecting and understanding hallucination.

5.4 Societal Impact

While this work primarily centers on the creation of a novel dataset, the potential societal impacts are significant. HaloQuest aims to provide a crucial tool for mitigating hallucination in VLMs, thereby improving their robustness and

reducing the likelihood of erroneous or misleading outputs. This has implications for real-world applications where safety and reliability are paramount, such as autonomous systems or medical image analysis. However, it is important to acknowledge that like any technology, datasets can be used for both beneficial and potentially harmful purposes. Bad actors could leverage datasets like HaloQuest to intentionally train models to generate misleading or deceptive content tailored to exploit model weaknesses. This fact underscores the importance of ongoing research into the detection and mitigation of such malicious use of AI systems.

6 Conclusion

This work has introduced HaloQuest, a novel VQA benchmark that leverages both real-world and synthetically generated images. HaloQuest's controlled image generation and questions designed to elicit specific hallucination types enable a more targeted analysis of hallucination triggers in VLMs. Experiments demonstrate that current state-of-the-art models struggle with HaloQuest, revealing a crucial disconnect between their capabilities and real-world reliability requirements. Importantly, fine-tuning VLMs on HaloQuest demonstrably reduces hallucination rates while maintaining performance on typical reasoning tasks.

HaloQuest highlights the potential of synthetic images in the development of robust multimodal AI. It addresses limitations present in traditional datasets, enabling the creation of richer and more varied visual scenarios. The dataset, coupled with an innovative machine-human-in-the-loop generation process, facilitates targeted investigation into VLM weaknesses.

Further, this work introduces an LLM-based Auto-Eval mechanism that facilitates open-ended and nuanced evaluation of VLM responses. This approach is a marked improvement over existing methods that often limit the model's expressive ability or are impractical for evaluating complex hallucinations.

HaloQuest stands as a valuable resource for the vision-and-language community. It provides both a challenging evaluation benchmark and a training dataset aimed at mitigating hallucination in VLMs. This work underscores the power of synthetic image generation and advanced evaluation techniques in driving the creation of more reliable and trustworthy multimodal AI systems.

Acknowledgements

The authors thank Ed Chi, Roopal Garg, Mohammad Norouzi, Kai-wei Chang, Shih-fu Chang, Junzhang Liu, Haoxuan You, Rui Sun and Chris Thomas for invaluable discussions. The authors would also like to thank Surge AI, in particular John Romeral and Paul Mains, for help with data collection.

References

1. Gpt-4v(ision) system card (2023), https://api.semanticscholar.org/CorpusID:263218031

2. Midjourney. https://midjourney.com/ (2023)
3. Alkaissi, H., Mcfarlane, S.: Artificial hallucinations in chatgpt: Implications in scientific writing. Cureus **15** (02 2023). https://doi.org/10.7759/cureus.35179
4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
5. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)
6. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Goldstein, J., Lavie, A., Lin, C.Y., Voss, C. (eds.) Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan (Jun 2005), https://aclanthology.org/W05-0909
7. Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., et al.: A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. arXiv preprint arXiv:2302.04023 (2023)
8. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. pp. 610–623 (2021)
9. Biten, A.F., Gómez, L., Karatzas, D.: Let there be a clock on the beach: Reducing object hallucination in image captioning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1381–1390 (2022)
10. Bitton-Guetta, N., Bitton, Y., Hessel, J., Schmidt, L., Elovici, Y., Stanovsky, G., Schwartz, R.: Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2616–2627 (2023)
11. Cui, C., Zhou, Y., Yang, X., Wu, S., Zhang, L., Zou, J., Yao, H.: Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. arXiv preprint arXiv:2311.03287 (2023)
12. Dai, W., Liu, Z., Ji, Z., Su, D., Fung, P.: Plausible may not be faithful: Probing object hallucination in vision-language pre-training. arXiv preprint arXiv:2210.07688 (2022)
13. Deng, J., Chan, G., Zhong, H., Lu, C.X.: See beyond seeing: Robust 3d object detection from point clouds via cross-modal hallucination. arXiv preprint arXiv:2309.17336 (2023)
14. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6904–6913 (2017)
15. Gunjal, A., Yin, J., Bas, E.: Detecting and preventing hallucinations in large vision language models. arXiv preprint arXiv:2308.06394 (2023)
16. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.: Survey of hallucination in natural language generation. ACM Computing Surveys **55**(12), 1–38 (2023)
17. Jiang, C., Ye, W., Dong, M., Jia, H., Xu, H., Yan, M., Zhang, J., Zhang, S.: Haleval: A universal and fine-grained hallucination evaluation framework for large vision language models. arXiv preprint arXiv:2402.15721 (2024)

18. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**, 32–73 (2017)
19. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. International Journal of Computer Vision **128**(7), 1956–1981 (2020)
20. Lee, S., Park, S.H., Jo, Y., Seo, M.: Volcano: mitigating multimodal hallucination through self-feedback guided revision. arXiv preprint arXiv:2311.07362 (2023)
21. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
22. Li, J., Cheng, X., Zhao, W.X., Nie, J.Y., Wen, J.R.: Halueval: A large-scale hallucination evaluation benchmark for large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 6449–6464 (2023)
23. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355 (2023)
24. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (Jul 2004), https://aclanthology.org/W04-1013
25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)
26. Liu, F., Guan, T., Li, Z., Chen, L., Yacoob, Y., Manocha, D., Zhou, T.: Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multimodality models. arXiv preprint arXiv:2310.14566 (2023)
27. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Aligning large multi-modal model with robust instruction tuning. arXiv preprint arXiv:2306.14565 (2023)
28. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253 (2024)
29. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744 (2023)
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
31. Lovenia, H., Dai, W., Cahyawijaya, S., Ji, Z., Fung, P.: Negative object presence evaluation (nope) to measure object hallucination in vision-language models. arXiv preprint arXiv:2310.05338 (2023)
32. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
33. Muhovič, J., Koporec, G., Perš, J.: Hallucinating hidden obstacles for unmanned surface vehicles using a compositional model (2023)
34. Pan, J., Sun, K., Ge, Y., Li, H., Duan, H., Wu, X., Zhang, R., Zhou, A., Qin, Z., Wang, Y., Dai, J., Qiao, Y., Li, H.: Journeydb: A benchmark for generative image understanding (2023)

35. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation (10 2002). https://doi.org/10.3115/1073083.1073135
36. Park, J.S., Xiao, X., Warnell, G., Yedidsion, H., Stone, P.: Learning perceptual hallucination for multi-robot navigation in narrow hallways. In: 2023 IEEE International Conference on Robotics and Automation (ICRA). pp. 10033–10039 (2023). https://doi.org/10.1109/ICRA48891.2023.10161327
37. Peng, D.: Langfun (Sep 2023), https://github.com/google/langfun
38. Qian, Y., Zhang, H., Yang, Y., Gan, Z.: How easy is it to fool your multimodal llms? an empirical analysis on deceptive prompts. arXiv preprint arXiv:2402.13220 (2024)
39. Rawte, V., Sheth, A., Das, A.: A survey of hallucination in large foundation models. arXiv preprint arXiv:2309.05922 (2023)
40. Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
41. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4035–4045 (2018)
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
43. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII. pp. 146–162. Springer (2022)
44. Sedgwick, P.: Pearson's correlation coefficient. BMJ **345**, e4483–e4483 (07 2012). https://doi.org/10.1136/bmj.e4483
45. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.Y., Wang, Y.X., Yang, Y., et al.: Aligning large multimodal models with factually augmented rlhf. arXiv preprint arXiv:2309.14525 (2023)
46. Umapathi, L.K., Pal, A., Sankarasubbu, M.: Med-halt: Medical domain hallucination test for large language models. arXiv preprint arXiv:2307.15343 (2023)
47. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
48. Wang, B., Wu, F., Han, X., Peng, J., Zhong, H., Zhang, P., Dong, X., Li, W., Li, W., Wang, J., et al.: Vigc: Visual instruction generation and correction. arXiv preprint arXiv:2308.12714 (2023)
49. Wang, H., Wu, W., Dou, Z., He, L., Yang, L.: Performance and exploration of chatgpt in medical examination, records and education in chinese: Pave the way for medical ai. International Journal of Medical Informatics **177**, 105173 (2023). https://doi.org/https://doi.org/10.1016/j.ijmedinf.2023.105173, https://www.sciencedirect.com/science/article/pii/S1386505623001910
50. Wang, J., Zhou, Y., Xu, G., Shi, P., Zhao, C., Xu, H., Ye, Q., Yan, M., Zhang, J., Zhu, J., et al.: Evaluation and analysis of hallucination in large vision-language models. arXiv preprint arXiv:2308.15126 (2023)
51. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: Ofa: Unifying architectures, tasks, and modalities through a simple

sequence-to-sequence learning framework. In: International Conference on Machine Learning. pp. 23318–23340. PMLR (2022)
52. Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., et al.: Image as a foreign language: Beit pretraining for all vision and vision-language tasks. arXiv preprint arXiv:2208.10442 (2022)
53. Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. arXiv:2210.14896 (2022)
54. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)
55. Ye, Q., Xu, H., Ye, J., Yan, M., Liu, H., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. arXiv preprint arXiv:2311.04257 (2023)
56. Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., Chen, E.: Woodpecker: Hallucination correction for multimodal large language models. arXiv preprint arXiv:2310.16045 (2023)
57. You, H., Sun, R., Wang, Z., Chen, L., Wang, G., Ayyubi, H.A., Chang, K.W., Chang, S.F.: Idealgpt: Iteratively decomposing vision and language reasoning via large language models. arXiv preprint arXiv:2305.14985 (2023)
58. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 **2**(3), 5 (2022)
59. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6720–6731 (2019)
60. Zhai, B., Yang, S., Zhao, X., Xu, C., Shen, S., Zhao, D., Keutzer, K., Li, M., Yan, T., Fan, X.: Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. arXiv preprint arXiv:2310.01779 (2023)
61. Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., Yao, H.: Analyzing and mitigating object hallucination in large vision-language models. arXiv preprint arXiv:2310.00754 (2023)
62. Zhu, D., Chen, J., Haydarov, K., Shen, X., Zhang, W., Elhoseiny, M.: Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. arXiv preprint arXiv:2303.06594 (2023)
63. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)