Mitigating Hallucinations in Large Vision-Language Models (LVLMs) via Language-Contrastive Decoding (LCD)

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) are an extension of Large Language Models (LLMs) that facilitate processing both image and text inputs, expanding AI capabilities. However, LVLMs struggle with object hallucinations due to their reliance on text cues and learned object co-occurrence biases. While most research quantifies these hallucinations, mitigation strategies are still lacking. Our study introduces a Language Contrastive Decoding (LCD) algorithm that adjusts LVLM outputs based on LLM distribution confidence levels, effectively reducing object hallucinations. We demonstrate the advantages of LCD in leading LVLMs, showing up to %4 improvement in POPE F1 scores and up to %36 reduction in CHAIR scores on the COCO validation set, while also improving captioning quality scores. Our method effectively improves LVLMs without needing complex post-processing or retraining, and is easily applicable to different models. Our findings highlight the potential of further exploration of LVLM-specific decoding algorithms for improved multimodal performance.

Figure 1: An illustration of LLM vs. LVLM token probabilities given an image and a text prefix mid-generation. According to the LLM, the word "dog" is much more likely to appear next. LCD dynamically contrasts these probabilities to mitigate language biases in LVLM outputs.

1 Introduction

Large Vision-Language Models (LVLMs) are a multimodal extension of Large Language Models (LLMs), transforming textual prompts and image inputs into text. However, they frequently produce object hallucinations, where absent objects are mentioned in the output (Yifan Li and Wen, 2023; Lovenia et al., 2023).

While hallucination-mitigation techniques in LLMs are actively researched, specific strategies for LVLMs are less developed. Current methods involve model-specific adjustments, additional training, or auxiliary models for post-hoc correction, and are often proven inefficient, costly, or limited by training data and model biases (Wang et al., 2023; Zhou et al., 2023; Gunjal et al., 2023; Yin et al., 2023). Conversely, LVLM hallucination evaluation has seen progress with object hallucination benchmarks like NOPE (Lovenia et al., 2023) and POPE (Yifan Li and Wen, 2023), and recent works that aim for more holistic LVLM hallucination evaluation such as FaithScore (Jing et al., 2023) and HallusionBench (Guan et al., 2023).

A key reason for LVLM hallucinations is their tendency to over-rely on linguistic information, as was first observed by Guan et al. (2023). Based on this insight, we propose to intervene in the LVLM decoding phase so that model outputs are less informed by language biases. Specifically, we propose to use Contrastive Decoding (Li et al., 2023a;

O'Brien and Lewis, 2023) to alter LVLM output probabilities with respect to the internal LLM's probabilities, guided by a dynamic weighting mechanism based on the LLM distribution's entropy.

Our experiments show that our proposed method, Language Contrastive Decoding (LCD), improves hallucination scores on POPE (Yifan Li and Wen, 2023) and CHAIR (Rohrbach et al., 2018) on InstructBLIP variants based on Vicuna and Flan-T5 (Dai et al., 2023), and LLAVA 1.5 (Liu et al., 2023). We asses LCD's overall generation quality by reporting captioning metrics and conducting a GPT4-V (OpenAI et al., 2023) assisted evaluation. LCD, as a decoding strategy, can be applied to other models without additional training or output modifications, emphasizing its utility for broader LVLM use.

The contributions of this paper are thus manifold. First, we introduce a novel decoding method tailored for LVLMs to mitigate object hallucinations. Next, we present a dynamic weighting strategy based on entropy which is applicable in various CD scenarios. Finally, we share our dataset and code, including a user-friendly LCD implementation via Huggingface, to encourage further research into LVLM-specific decoding strategies, a promising avenue for future research.

2 Motivation and Background

The integration of vision capabilities into Large Language Models has led to the development of Large Vision-Language Models, merging LLMs' textual understanding with vision-text encoders. This trend towards multimodal systems is exemplified in commercial platforms such as GPT4-V (OpenAI et al., 2023) and Google's Gemini (Team et al., 2023).

Large Vision-Language Models combine LLMs and vision-text encoders to generate text from textual prompts and visual inputs. An LVLM generally comprises three main components: a vision-text encoder like CLIP (Radford et al., 2021), an LLM such as LLAMA (Touvron et al., 2023) or Flan-T5 (Chung et al., 2022), and a cross-modal alignment module linking the vision-text encoder output with the LLM.

Initially, LVLMs were fine-tuned for specific tasks (Li et al., 2022; Wang et al., 2022). However, advancements in LLMs have led to a shift towards general-purpose, instruction-tuned LVLMs. These models are designed to handle a wide range of tasks based on instructions, making them more versatile. Despite these advancements, LVLMs grapple with hallucinations of different types.

LVLMs Hallucinations and their Mitigation Hallucinations in LVLMs, particularly object hallucinations where nonexistent entities are mentioned, are often attributed to LVLMs' reliance on spurious correlations and language biases, as demonstrated by Li et al. (2023b) and Zhou et al. (2023). Moreover, Guan et al. (2023) highlight LVLMs' tendency to prioritize language over visual data, leading to hallucinations.

Mitigation strategies proposed by Gunjal et al. (2023) and Wang et al. (2023) involve further model training with augmented datasets or reward models. Zhou et al. (2023); Yin et al. (2023) developed auxiliary models to correct outputs post-generation. These solutions often require dataset-specific work or additional model training, potentially leading to overfitting or new biases, and are not easily transferable across LVLMs. In this work we aim to address these shortcomings with a more general, modular approach.

3 Language Contrastive Decoding (LCD)

Before presenting LCD, we briefly introduce the essentials of decoding in LVLMs 3.1, followed by our formal proposal 3.2 and research hypothesis 3.3.

3.1 Decoding Techniques and Contrastive Decoding: Essential Preliminaries

Decoding in auto-regressive generative models is the stage that transforms an input representation into a sequence of output tokens. In LVLMs, this process involves a model M, an image I, a textual prompt X, and a particular timestamp t during generation. It can be described as a series of selections from the model's probability distribution, producing a token sequence T, as formalized in eq. (1).

$$T_t \sim P(\cdot|I, X, T_{<t}; M) \tag{1}$$

Greedy decoding, selecting the most probable token at each step (or the top k tokens in a beam search with k beams), is the simplest approach. However, high likelihood sequences do not necessarily align with human preferences, leading to the "likelihood trap" (Zhang et al., 2021). This has led to the use of sampling-based methods, such as top-k sampling, nucleus sampling (Holtzman et al., 2020), and locally typical sampling (Meister et al., 2023), which either truncate the set of candidate tokens or adjust the model's distribution, e.g. through temperature scaling.

Contrastive Decoding (CD) has been introduced for LLMs as a method to penalize the outputs of an expert model with those from a less powerful model (Li et al., 2023a). CD can be applied to any two probability distributions with the same support and has been adapted as a sampling strategy, improving various text generation tasks (O'Brien and Lewis, 2023; Chuang et al., 2023; Sennrich et al., 2024). CD uses both truncation and reshaping of probability distributions. The truncation phase ("adaptive plausibility") is described by eq. (2), where α is a hyper-parameter, $\mathcal{V}$ and $\mathcal{V}t'$ are the original and truncated token vocabularies at time t, and P is the conditional distribution on the prefix $T_{<t}$.

$$\mathcal{V}t' = \{v \in \mathcal{V} : P(v|T_{<t}) \geq \alpha \max_{w} P(w|T_{<t})\} \tag{2}$$

Finally, the formula for CD, as suggested by O'Brien and Lewis (2023), given here generally for two conditional distributions P and P' on variable x with the same support, conditioned on X is presented in eq. (3).

$$CDt(x, X, P, P') = \begin{cases} (1+\beta)\log P(x|X) - \beta \log P'(x|X), & \text{if } x \in Vt' \\ -\infty, & \text{otherwise} \end{cases} \tag{3}$$

β is a fixed weight hyper-parameter. Our proposed method, detailed shortly, alters CD by introducing an entropy-based dynamic weighting scheme.

3.2 Proposed Method

Our intuition, based on previous findings by (Guan et al., 2023; Rohrbach et al., 2018; Yifan Li and Wen, 2023), is that an LVLM can be "misled" by its constituent LLM during the generation process.

Consider for example an LVLM that is describing an image (see illustration 1). Mid-generation, given the text "An image of a man walking his," it may predict "dog" due to language biases, even if it is a bear shown that is actually shown. A 'plain' LLM, without seeing the image, reinforces these biases by highly rating "dog". Our method builds on this insight to guide an LVLM towards more accurate predictions using Contrastive Decoding.

Our method operates as follows: At each generation step t, for each token x, we first determine the next-token probabilities from the LVLM, P_{LVLM}, based on the current token sequence $T_{<t}$, text X, and image I. We then obtain a second distribution, P_{LLM}, by inputting all data except the image into the LLM. The LLM's conditional entropy H_{LLM} informs the dynamic weight as per eq.4. We then adjust token x's logits using the LCD formula in eq. 5.

$$\beta_t = \frac{\beta}{\mathrm{H}_{LLM}(x|X, T_{<t})} \tag{4}$$

$$\begin{aligned} LCD_t(x, T_{<t}, I, P_{LVLM}, P_{LLM}) = \\ (1+\beta_t)\log P_{LVLM}(x|I, X, T_{<t}) \\ - \beta_t \log P_{LLM}(x|X, T_{<t}) \end{aligned} \tag{5}$$

We use the LCD logits for sampling, with temperature set to 1.0 in all experiments unless stated otherwise.

3.3 Research Hypothesis

Our hypothesis is that contrasting LVLM outputs with LLM outputs conditioned only on the textual data, can mitigate language biases, therefore reducing hallucinations in LVLMs.

4 Experiments and Results

We set out to assess the effect of LCD on object hallucinations in LVLM outputs against popular decoding settings. Additionally, we verify that LCD does not degrade output quality. To this end, we asses LCD on the POPE benchmark (Yifan Li and Wen, 2023), and on an image detailed-description task where we report hallucination and captioning metrics and conduct a GPT4-V assisted evaluation.

Polling-based Object-Probing Evaluation POPE consists of object-presence binary questions on 500 COCO dataset images (Lin et al., 2015), with questions equally divided between present and absent objects. Absent objects are chosen based on three criteria: *random*, *popular* (common in COCO), and *adversarial* (commonly co-occurring with present objects). POPE's drawback is its one-word response structure, which limits the influence of decoding strategies and does not evaluate open-ended generation capabilities.

Model	Method	METEOR↑	WMD↑	ROUGE_L↑	Acc↑	Det↑	CHAIRs↓	CHAIRi↓
InstructBLIP$_F$	Baseline	.157	.367	.161	4.92	4.02	.662	.146
	LCD	**.159**	**.370**	**.168**	**5.4**	4.01	**.566**	**.131**
InstructBLIP$_V$	Baseline	.178	.423	.291	3.7	3.51	.274	.126
	LCD	**.199**	**.48**	**.38**	**4.59**	**3.83**	**.174**	**.107**
LLAVA 1.5	Baseline	.163	**.357**	.169	4.77	4.56	.672	.182
	LCD	**.171**	.352	**.181**	**5.39**	4.54	**.610**	**.161**

Table 1: Image Description results. F and V stand for the Flan-T5 and Vicuna. Acc and Det are mean GPT4-V scores for Accuracy and Detailedness. METEOR, WMD and ROUGE_L are popular captioning metrics (Kusner et al., 2015; Banerjee and Lavie, 2005; Lin, 2004).

Image Detailed-Descriptions To complement POPE, we introduce a long-form text generation task, inspired by findings from Zhou et al. (2023), that more extensive context increases hallucinations. The task uses the same COCO images as POPE, for which we generate detailed descriptions. The prompts we use are detailed in appendix A.1. This task tests LCD in a scenario where the propensity to hallucinate is high, and its longer outputs facilitate qualitative evaluation.

Baselines and Metrics For POPE, we use sampling as the baseline and report F1 scores.[1] For the descriptions task, we use the popular nucleus sampling algorithm[2] and report CHAIR metrics (Rohrbach et al., 2018). To assess description quality, we use captioning metrics against COCO's gold captions, which serve as an approximation considering length differences. Additionally, following Yin et al. (2023), we use GPT4-V to evaluate the descriptions for Detailedness and Accuracy (details provided in Appendix A.1).

Models We conduct our experiments with leading LVLMs: two versions of the InstructBLIP model (with Flan-T5 and Vicuna LLMs) and LLAVA 1.5. The complete experimental details, such as exact model variants and generation hyper-parameters, are given in the appendix.

5 Results and Discussion

In POPE, LCD improves F1 scores across 8 of 9 configurations (table 2). In the image description task (table 1), it significantly reduces hallucinations at both sentence and instance levels in all three models. However, CHAIR numbers are still high (especially in the Vicuna based models, InstructBLIP$_\text{V}$ and LLAVA), demonstrating the prevalence of object hallucinations in long form LVLM outputs. Notably, InstructBLIP$_\text{V}$ outperforms LLAVA 1.5 in most assessed metrics, in spite of being considered a weaker model. LCD shows particular effectiveness in InstructBLIP models probably due to their LLMs being frozen during training, which lends them a stronger language bias. In terms of overall generation quality as reflected by the captioning metrics, LCD is better in all cases but one (WMD in LLAVA 1.5, where the reduction is ~%1). In the GPT4-V eval, LCD improves Accuracy in all cases, and is as detailed as the baseline, suggesting it reduces hallucinations but does not increase the granularity of the descriptions.

POPE	Model	Baseline F1	LCD F1
Random	InstructBLIP$_V$	83.95	**87.55**
Popular		82.80	**84.34**
Adversarial		80.25	**81.64**
Random	InstructBLIP$_F$	84.05	**84.27**
Popular		80.74	**82.81**
Adversarial		78.87	**80.69**
Random	LLAVA 1.5	**84.17**	83.76
Popular		83.10	**83.47**
Adversarial		81.34	**81.62**

Table 2: POPE results for different models and methods.

6 Conclusion

In this paper we present Language Contrastive Decoding, a novel method to reduce hallucinations in LVLMs. By dynamically adjusting output probabilities using the LVLM's internal LLM, LCD significantly improves hallucination metrics across different LVLM architectures, enhancing the quality and reliability of generated content without necessitating retraining or auxiliary models and post-processing. This work highlights the potential of specialized decoding strategies in enhancing multimodal AI models and lays the groundwork for further exploration into more sophisticated LVLM decoding methods.

[1] Complete POPE results are in the appendix, table 4

[2] We find that nucleus-sampling gives better results than vanilla sampling (see table 3 in the appendix for ablations)

7 Limitations

Firstly, while LCD shows promise in reducing hallucinations, it only targets hallucinations caused by language biases, but hallucinations can arise from other sources. For instance, previous work has shown that some hallucinations are caused by poor visual understanding (Guan et al., 2023). We believe LCD can be used as a platform to craft LVLM-specific decoding algorithms that would mitigate hallucinations stemming from different factors, and leave this pursuit for future work.

Secondly, our evaluation method primarily addresses object hallucinations, which are only one form of hallucination that LVLMs may exhibit. Preliminary results signal that LCD mitigates more complex manifestations of language-induced hallucinations as assessed by recent benchmarks such as FAITHSCORE (Jing et al., 2023) and HallusionBench (Guan et al., 2023), but further work is required to establish this.

Moreover, LCD relies on current LVLM architectures that combine an LLM and a text-vision encoder, and requires access to an LLM that emits output probabilities on the same set of tokens as the LVLM. It is possible that the future generation of multimodal AI systems will have a different architecture that will make LCD obsolete. Additionally, LCD requires an LLM forward pass for each LVLM decoding step. The added latency could be mitigated with efficient inference techniques, and also by using a smaller LLM as the contrasting model. The effectiveness of LCD in this scenario is left for future work.

Finally, there are ethical considerations related to the mitigation of hallucinations in LVLMs. As these models become more reliable, it is crucial to continue evaluating the potential impacts of their use, ensuring they do not perpetuate or exacerbate biases present in their training data. LCD indeed mitigates some biases, but it is important to keep in mind that it might amplify other biases, unknown to us. Responsible deployment of these models requires ongoing vigilance and a commitment to transparency and fairness.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination visual illusion in large vision-language models.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2023. Detecting and preventing hallucinations in large vision language models.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration.

Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. 2023. Faithscore: Evaluating hallucinations in large vision-language models.

M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger. 2015. From word embeddings to document distances. In *ICML*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023a. Contrastive decoding: Open-ended text generation as optimization.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally typical sampling.

Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. 2024. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Martin Chadwick, Gaurav Singh Tomar, Xavier Garcia, Evan Senter, Emanuel Taropa, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Yujing Zhang, Ravi Addanki, Antoine Miech, Annie Louis, Laurent El Shafey, Denis Teplyashin, Geoff Brown, Elliot Catt, Nithya Attaluri, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaly Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, Hanzhao Lin, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yong Cheng, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Ro-

driguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, YaGuang Li, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Gamaleldin Elsayed, Ed Chi, Mahdis Mahdieh, Ian Tenney, Nan Hua, Ivan Petrychenko, Patrick Kane, Dylan Scandinaro, Rishub Jain, Jonathan Uesato, Romina Datta, Adam Sadovsky, Oskar Bunyan, Dominik Rabiej, Shimu Wu, John Zhang, Gautam Vasudevan, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Betty Chan, Pam G Rabinovitch, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Sahitya Potluri, Jane Park, Elnaz Davoodi, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Chris Gorgolewski, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Paul Suganthan, Evan Palmer, Geoffrey Irving, Edward Loper, Manaal Faruqui, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Michael Fink, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marin Georgiev, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Minnie Lui, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Lam Nguyen Thiet, Daniel Andor, Pedro Valenzuela, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Sarmishta Velury, Sebastian Krause, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Tejasi Latkar, Mingyang Zhang, Quoc Le, Elena Allica Abellan, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Sid Lall, Ken Franko, Egor Filonov, Anna Bulanova, Rémi Leblond, Vikas Yadav, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Hao Zhou, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Jeremiah Liu, Mark Omernick, Colton Bishop, Chintu Kumar, Rachel Sterneck, Ryan Foley, Rohan Jain, Swaroop Mishra, Jiawei Xia, Taylor Bos, Geoffrey Cideron, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Petru Gurita, Hila Noga, Premal Shah, Daniel J. Mankowitz, Alex Polozov, Nate Kushman, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Anhad Mohananey, Matthieu Geist, Sidharth Mudgal, Sertan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Quan Yuan, Sumit Bagri, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Aliaksei Severyn, Jonathan Lai, Kathy Wu, Heng-Tze Cheng, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Mark Geller, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Andrei Sozanschi, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Abhimanyu Goyal, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Sabaer Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Tao Zhu, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Dustin Tran, Yeqing Li, Nir Levine, Ariel Stolovich, Norbert Kalb, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Balaji Lakshminarayanan, Charlie Deck, Shyam Upadhyay, Hyo Lee, Mike Dusenberry, Zonglin Li, Xuezhi Wang, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Summer Yue, Sho Arora, Eric Malmi, Daniil Mirylenka, Qijun Tan, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Steven Zheng, Francesco Pongetti, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Ragha Kotikalapudi, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Chenkai Kuang, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Pei Sun, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Ishita Dasgupta, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Yuan Liu, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Ce-

sare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Ivo Penchev, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Adam Kurzrok, Lynette Webb, Sahil Dua, Dong Li, Preethi Lahoti, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Taylan Bilal, Evgenii Eltyshev, Daniel Balle, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Adams Yu, Christof Angermueller, Xiaowei Li, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Kevin Brooks, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Komal Jalan, Dinghua Li, Ginger Perng, Blake Hechtman, Parker Schuh, Milad Nasr, Mia Chen, Kieran Milan, Vladimir Mikulik, Trevor Strohman, Juliana Franco, Tim Green, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A family of highly capable multimodal models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2023. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites.

Kun Zhou Jinpeng Wang Wayne Xin Zhao Yifan Li, Yifan Du and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. 2021. Trading off diversity and quality in natural language generation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 25–33, Online. Association for Computational Linguistics.

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

A Appendix

A.1 Detailed Experimental Setup

For POPE and the descriptions experiment, we use the following LCD parameters $\beta = 3.0$, $\alpha = 0.1$. We set the temperature to 0.5 in POPE and 1.0 in the descriptions experiment. We limit the descriptions length to 250 tokens in all models we tested. We don't tune any of these parameters. The prompt we use for the descriptions experiment is *"Describe this image in detail:"*. The models we use have the following Huggingface identifiers:

- Salesforce/instructblip-vicuna-7b
- Salesforce/instructblip-flan-t5-xl
- llava-hf/llava-1.5-7b-hf

GPT4-V Assisted Evaluation We follow the evaluation protocol given in Yin et al. (2023), where an image and two descriptions are given to the model, formatted with the prompt in figure 2. The model outputs scores in two dimensions: Accuracy and Detailedness. We use the *gpt-4-vision-preview* model on February 2024.

```
prompt = lambda A, B: f"""
You are required to score the performance of two AI assistants in describing a given image. You should pay extra
attention to the hallucination, which refers to the part of descriptions that are inconsistent with the image content,
such as claiming the existence of something not present in the image or describing incorrectly in terms of the counts,
positions, or colors of objects in the image. Note that the descriptions may be accompanied by bounding boxes,
indicating the position of objects in the image, which are represented as [x1, y1, x2, y2] with floating numbers ranging
from 0 to 1. These values correspond to the top left x1, top left y1, bottom right x2, and bottom right y2.
Please rate the responses of the assistants on a scale of 1 to 10, where a higher score indicates better performance,
according to the following criteria:
1: Accuracy: whether the response is accurate with respect to the image content. Responses with fewer hallucinations
should be given higher scores.
2: Detailedness: whether the response is rich in necessary details. Note that hallucinated descriptions should not count
as necessary details.
Please output a single line for each criterion, containing only two values indicating the scores for Assistant 1 and 2,
respectively. The two scores are separated by a space. Following the scores, please provide an explanation of your
evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not
affect your judgment.

[Assistant 1]
{A}
[End of Assistant 1]

[Assistant 2]
{B}
[End of Assistant 2]

Output format:

Accuracy:
Scores of the two answers:
Reason:

Detailedness:
Scores of the two answers:
Reason:
"""
```

Figure 2: Prompt used to evaluate descriptions with GPT4-V, taken from Yin et al. (2023)

A.2 Detailed Experimental Results

Model	Method	METEOR↑	WMD↑	ROUGE_L↑	CHAIRs↓	CHAIRi↓
InstructBLIP$_F$	Baseline	0.151	0.361	0.156	0.666	0.174
	Baseline$_N$	0.157	0.367	0.161	0.662	0.146
	LCD$_{-dw}$	0.159	0.364	0.163	0.594	0.133
	LCD	**0.163**	**0.370**	**0.168**	**0.566**	**0.131**
InstructBLIP$_V$	Baseline	0.171	0.408	0.274	0.308	0.138
	Baseline$_N$	0.178	0.423	0.291	0.274	0.126
	LCD$_{-dw}$	**0.202**	0.474	0.366	0.23	0.116
	LCD	0.199	**0.48**	**0.38**	**0.174**	**0.107**
LLAVA 1.5	Baseline	0.160	0.353	0.167	0.632	0.183
	Baseline$_N$	0.163	**0.357**	0.169	0.672	0.182
	LCD$_{-dw}$	0.169	0.352	0.179	0.624	**0.157**
	LCD	**0.171**	0.352	**0.181**	**0.610**	0.161

Table 3: Image Description ablations. *-dw* is an LCD variant without dynamic weighting, with $\beta = 0.5$. Baseline$_{\mathrm{N}}$ is using nucleus sampling with $p = 0.95$, Baseline is vanilla sampling.

POPE	method	model	accuracy	precision	recall	f1	yes_ratio
random	Baseline	InstructBLIP Vicuna	84.90%	89.57%	79.00%	83.95%	44.10%
random	LCD	InstructBLIP Vicuna	87.53%	87.43%	87.67%	87.55%	50.13%
popular	Baseline	InstructBLIP Vicuna	83.30%	85.35%	80.40%	82.80%	47.10%
popular	LCD	InstructBLIP Vicuna	83.73%	81.31%	87.60%	84.34%	53.87%
adversarial	Baseline	InstructBLIP Vicuna	80.23%	80.17%	80.33%	80.25%	50.10%
adversarial	LCD	InstructBLIP Vicuna	80.27%	76.33%	87.73%	81.64%	57.47%
random	Baseline	InstructBLIP FlanT5	85.63%	94.43%	75.73%	84.05%	40.10%
random	LCD	InstructBLIP FlanT5	86.03%	96.47%	74.80%	84.27%	38.77%
popular	Baseline	InstructBLIP FlanT5	82.07%	87.17%	75.20%	80.74%	43.13%
popular	LCD	InstructBLIP FlanT5	84.43%	92.44%	75.00%	82.81%	40.57%
adversarial	Baseline	InstructBLIP FlanT5	79.83%	82.83%	75.27%	78.87%	45.43%
adversarial	LCD	InstructBLIP FlanT5	82.03%	87.22%	75.07%	80.69%	43.03%
random	Baseline	LLAVA 1.5	85.87%	95.67%	75.13%	84.17%	39.27%
random	LCD	LLAVA 1.5	85.73%	97.18%	73.60%	83.76%	37.87%
popular	Baseline	LLAVA 1.5	84.80%	93.57%	74.73%	83.10%	39.93%
popular	LCD	LLAVA 1.5	85.40%	96.17%	73.73%	83.47%	38.33%
adversarial	Baseline	LLAVA 1.5	82.77%	88.67%	75.13%	81.34%	42.37%
adversarial	LCD	LLAVA 1.5	83.33%	90.98%	74.00%	81.62%	40.67%

Table 4: Complete POPE results.