SEEING IS BELIEVING: MITIGATING HALLUCINATION IN LARGE VISION-LANGUAGE MODELS VIA CLIP-GUIDED DECODING

Ailin Deng, Zhirui Chen & Bryan Hooi
School of Computing
National University of Singapore
{ailin,zhiruichen}@u.nus.edu bhooi@comp.nus.edu.sg

ABSTRACT

Large Vision-Language Models (LVLMs) are susceptible to object hallucinations, an issue in which their generated text contains non-existent objects, greatly limiting their reliability and practicality. Current approaches often rely on the model's token likelihoods or other internal information, instruction tuning on additional datasets, or incorporating complex external tools. We first perform empirical analysis on sentence-level LVLM hallucination, finding that CLIP similarity to the image acts as a stronger and more robust indicator of hallucination compared to token likelihoods. Motivated by this, we introduce our CLIP-Guided Decoding (CGD) approach[1], a straightforward but effective training-free approach to reduce object hallucination at decoding time. CGD uses CLIP to guide the model's decoding process by enhancing visual grounding of generated text with the image. Experiments demonstrate that CGD effectively mitigates object hallucination across multiple LVLM families while preserving the utility of text generation.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) have shown impressive visual reasoning capabilities, representing an important milestone toward agents that can operate autonomously in our visual world (Achiam et al., 2023; Liu et al., 2023b; Dai et al.; Xi et al., 2023). However, object hallucination, in which the model produces inaccurate descriptions featuring non-existent objects, greatly limit the model's reliability and practical utility (Rohrbach et al., 2018; Wang et al., 2023c; Gunjal et al., 2023; Zhou et al., 2023). Hallucinations can easily mislead users, particularly when accompanied by overconfidence (Xiong et al., 2023). This is especially a concern for safety-critical applications such as robotics (Brohan et al., 2023) and medical image analysis (Thawkar et al., 2023), as well as for human-AI interaction settings where models should behave in a predictable manner that is well aligned with human expectations.

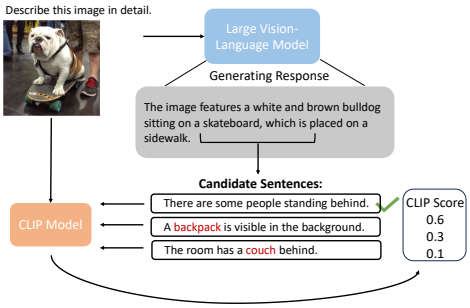


Figure 1: Intuition of our method: candidate sentences with higher CLIP similarity to the image are less likely to be hallucinated, and hence selected during the decoding process. Hallucinated text is colored red.

Intuitively, object hallucination can be viewed through a lens of human-AI misalignment. Consider how humans describe images: generally, we mentally 'anchor' or compare our descriptions to objects in the image, making hallucination unlikely. In contrast, LVLMs generate text based on token likelihoods without explicitly 'anchoring' or comparing to the image in this manner, making hallucination more likely. This gap between how humans and AI operate makes errors made by AI more surprising and unpredictable, which is detrimental in settings involving human-AI interaction.

[1]Codes are available at https://github.com/d-ailin/CGD

While LVLMs do rely on self-attention, the attention weight for a newly generated token is spread over both the image tokens, and the model's own previously generated text tokens, which can lead to a diminishing emphasis of the image tokens as caption length increases. Along these lines, we observe empirically that multiple LVLM architectures show a consistent increasing trend in their hallucination probability as a function of sentence index in the caption.

Recent efforts (Liu et al., 2023a; Wang et al., 2023a) have introduced approaches to mitigate hallucinations through instruction tuning. However, these methods come with substantial additional costs, including an annotation budget to acquire extra instruction data (Liu et al., 2023a). Some studies advocate leveraging internal information from models, such as likelihood scores (Zhou et al., 2023) and internal hidden states (Huang et al., 2023; Leng et al., 2023). Nevertheless, relying solely on a model's internal information may be insufficient (Manakul et al., 2023) and potentially unreliable, given the known overconfidence issues in neural networks (Kadavath et al., 2022; Xiong et al., 2023). Alternatively, incorporating external tools or knowledge has been suggested (Yin et al., 2023), but this approach often involves separate, occasionally intricate modules, accompanied by a heavy engineering burden (Yin et al., 2023).

In contrast with these approaches, we mitigate hallucination by directly comparing the LVLM's generated text to the image to ensure good correspondence between them, improving alignment to how humans perform the task. As illustrated in Figure 1, we introduce CLIP-Guided Decoding (CGD), a straightforward but effective training-free approach that utilizes CLIP as an external guide to alleviate hallucination during decoding.

2 PRELIMINARIES

Notations. In this context, we denote input $\boldsymbol{x} = (\boldsymbol{x}_{\text{img}}, \boldsymbol{x}_{\text{text}})$ including an input image $\boldsymbol{x}_{\text{img}}$ and a text input $\boldsymbol{x}_{\text{text}}$, e.g. 'Describe this image in detail'. In addition, we represent the generated response $\boldsymbol{y}$ in the form of L sequential sentences as $\boldsymbol{y} \coloneqq (\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L)$, where a sentence $\boldsymbol{s}_i$ is the i-th sentence and consists of l_i tokens: $\boldsymbol{s}_i \coloneqq (z_1^{(i)}, \ldots, z_{l_i}^{(i)})$ for $i \in [L]$ and $[L] \coloneqq \{1, \ldots, L\}$. Additionally, we define $H(\boldsymbol{s}_i) \in \{0, 1\}$ to output the hallucination label given a sentence $\boldsymbol{s}_i$. That is, $H(\boldsymbol{s}_i) = 1$ if the sentence $\boldsymbol{s}_i$ contains an incorrect object for the image $\boldsymbol{x}_{\text{img}}$. Since we focus on hallucination detection at the sentence level, we first reformulate the generative process as sentence generation.

Sentence Likelihood. Given a Large Vision-Language Model (LVLM) model with parameters $\boldsymbol{\theta}$, an image $\boldsymbol{x}_{\text{img}}$ and text input $\boldsymbol{x}_{\text{text}}$, the model generates a response $\boldsymbol{y}$ by conditional sentence generation in an auto-regressive manner (Ranzato et al., 2015): $\log p_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{x}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_L \mid \boldsymbol{x}) = \sum_{i=1}^{L} \log p_{\boldsymbol{\theta}}(\boldsymbol{s}_i \mid \boldsymbol{x}, \boldsymbol{s}_{<i})$, where $\boldsymbol{s}_{<i}$ are the sentences before i-th sentence in the response $\boldsymbol{y}$. Further, sentence generation is a process of conditional token generation: $\log p_{\boldsymbol{\theta}}(\boldsymbol{s}_i \mid \boldsymbol{x}, \boldsymbol{s}_{<i}) = \log \sum_{j=1}^{l_i} p_{\boldsymbol{\theta}}(z_j^{(i)} \mid \boldsymbol{x}, \boldsymbol{s}_{<i}, z_{<j}^{(i)})$, where $z_{<j}^{(i)}$ are the tokens before position j in sentence $\boldsymbol{s}_i$. Though $\log p_\theta(\boldsymbol{y} \mid \boldsymbol{x})$ is statistically meaningful, it has been shown to be biased with sequence length, i.e., models tend to overly favor shorter sentences (Wu et al., 2016). Length-normalized likelihood is an alternative, where the normalization can be either over the response or the sentence:

$$f_{\boldsymbol{\theta}}(\boldsymbol{y}) \coloneqq \frac{1}{\sum_{i=1}^{L} l_i} \sum_{i=1}^{L} \log p_{\boldsymbol{\theta}}(\boldsymbol{s}_i \mid \boldsymbol{x}, \boldsymbol{s}_{<i}), \quad g_{\boldsymbol{\theta}}(\boldsymbol{s}_i) \coloneqq \frac{1}{l_i} \log p_{\boldsymbol{\theta}}(\boldsymbol{s}_i \mid \boldsymbol{x}, \boldsymbol{s}_{<i}), \tag{1}$$

where $f_{\boldsymbol{\theta}}(\boldsymbol{y})$ is the length normalized likelihood of a response $\boldsymbol{y}$, and $g_{\boldsymbol{\theta}}(\boldsymbol{s}_i)$ is the sentence likelihood, the length normalized likelihood of the i-th sentence $\boldsymbol{s}_i$ conditioning on the previously sentences.

CLIP. Typically, an image-text contrastive model (Radford et al., 2021), e.g. CLIP model parameterized with $\boldsymbol{\phi}$, consists of image and text feature extractors: $f_{\boldsymbol{\phi}_{\text{img}}}$ and $f_{\boldsymbol{\phi}_{\text{text}}}$. Given an image $\boldsymbol{x}_{\text{img}}$ and a sentence $\boldsymbol{s}$, we obtain the CLIPScore as cosine similarity between the normalized image feature and text feature:

$$f_{\boldsymbol{\phi}}(\boldsymbol{x}_{\text{img}}, \boldsymbol{s}) \coloneqq \cos(f_{\boldsymbol{\phi}_{\text{img}}}(\boldsymbol{x}_{\text{img}}), f_{\boldsymbol{\phi}_{\text{text}}}(\boldsymbol{s})). \tag{2}$$

CLIPScore provides a quantitative metric to reflect how well a textual description is associated with the image and is widely applied in image-text evaluation (Hessel et al., 2021; Betker et al., 2023).

We include the hallucination analysis by comparing sentence likelihood and CLIPScore in Appendix B. The main observations include: **(1)** Later sentences are more prone to hallucinations; **(2)** Poor predictive performance of sentence likelihood; **(3)** there exists likelihood gap of sentence likelihood across datasets; **(4)** CLIP scores is a more effective and robust indicator of hallucination detection compared to sentence likelihood.

3 CLIP-GUIDED DECODING

Given the effectiveness of CLIPScore in detecting hallucination, we propose CLIP-Guided Decoding (CGD) to reduce hallucination by using CLIP as vision-language guidance to prefer visually grounded content during generation. The algorithm consists of two parts: *Reliability Scoring*, which designs a scoring function aiming to assign higher scores to candidate responses which are less likely to be hallucinated, and *Guided Sentence Generation*, which generates responses based on this scoring function. We decode in a similar way to beam search, but at the sentence level: this allows us to apply CLIP scoring on full sentences instead of incomplete words or phrases which would be present when decoding at the token level.

Reliability Scoring. Given a step t at decoding time and a candidate response $\mathbf{c}$ containing sequential sentences $(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_t)$, we define the reliability score as follows:

$$F(\mathbf{c}) \coloneqq (1-\alpha) f_\theta(\mathbf{c}) + \alpha \frac{1}{t} \sum_i^t f_\phi(\boldsymbol{x}_{\text{img}}, \boldsymbol{s}_i), \tag{3}$$

where $\alpha \in [0, 1]$ is hyperparameter to weigh the normalized likelihood $f_\theta(\mathbf{c})$ from the LVLM model, and the CLIP guidance $\sum_i^t f_\phi(\boldsymbol{x}_{\text{img}}, \boldsymbol{s}_i)$. Mixing between likelihoods and CLIP guidance gives users flexibility to control the strength of CLIP guidance. When $\alpha = 0$, CLIP guidance has no effect and the scoring function only considers the likelihood scores from the model. $\alpha = 1$ indicates the preference for higher CLIPScore.

Guided Sentence Generation. Next, we generate responses, guided by the scoring function. An important goal is to *maintain the generation quality and fluency* of the LVLM by preserving its sentence sampling process. To mitigate hallucinations, we use the reliability score function to prioritize responses that are *well-grounded* to the image.

Specifically, for every step t, we maintain a set of candidates with maximum cardinality N as $\mathcal{C}_t \coloneqq \{\mathbf{c}_1^t, \mathbf{c}_2^t, \ldots, \mathbf{c}_N^t\}$, where each candidate $\mathbf{c}_j^t \coloneqq (\boldsymbol{s}_{j,1}^t, \boldsymbol{s}_{j,2}^t, \ldots, \boldsymbol{s}_{j,t}^t)$ represents the first t sequences that are generated for $j \in [N]$.

Given a candidate set $\mathcal{C}_t$ in step t, the candidate set $\mathcal{C}_{t+1}$ is generated as follows. Firstly, for every $\mathbf{c}_j^t \in \mathcal{C}_t$, we independently sample its next sentence M times, following the conditional distribution parameterized by LVLM model parameter $\boldsymbol{\theta}$ (see Algorithm 1), from which we obtain a new candidate set $\mathcal{C}'_{t+1}$. Then, to avoid the cardinality of the candidate set exponentially increasing, we keep the top N candidates in $\mathcal{C}'_{t+1}$ based on the sentence-level hallucination scoring, i.e., $\mathcal{C}_{t+1} \coloneqq \underset{\mathbf{c} \in \mathcal{C}'_{t+1}}{\text{argtop}N} \{F(\mathbf{c})\}$, where argtop$N$ returns the top N members of the set in the subscript, with the highest values of the function (F). We repeat the procedure until we reach EOF for all candidates, e.g., reaching the maximum output length. Finally, we use the highest-scoring candidate $\mathbf{c}^*$ from $\mathcal{C}_t$ as output. We summarize CLIP-Guided Decoding in Algorithm 1.

4 EXPERIMENTS

Tasks and Datasets. Our evaluation tasks include hallucination evaluation and open-ended VQA, where hallucination evaluation is our main focus, and open-ended VQA allows us to gain insight into the method's effect in a wider array of settings. The hallucination evaluation includes COCO (Lin et al., 2014) (using Karpathy Test split (Karpathy & Fei-Fei, 2015)), NoCaps (near-domain) and Nocaps (out-of-domain) (Agrawal et al., 2019), where Nocaps (out-of-domain) contains the out-of-domain data regarding COCO objects. We randomly select 500 samples from each set for each run and prompt the LVLMs with "Describe this image in detail". For open-ended VQA, we use the

Table 1: Hallucination evaluation on COCO Karpathy Test Split (Karpathy & Fei-Fei, 2015) and NoCaps Validation Set (Near-Domain and Out-of-Domain) (Agrawal et al., 2019). Please see the case study in Table 6 in Appendix.

		InstructBLIP			mPLUG-Owl2			LLaVA-1.5		
		$C_S \downarrow$	$C_I \downarrow$	Avg. Len	$C_S \downarrow$	$C_I \downarrow$	Avg. Len	$C_S \downarrow$	$C_I \downarrow$	Avg. Len
COCO	Greedy	57.9	17.1	102.7	52.7	16.0	89.4	44.7	13.1	80.1
	Nucleus	56.1	17.0	98.3	51.9	15.6	89.0	43.3	13.1	80.1
	TopK	55.8	16.9	97.3	53.1	15.9	89.2	44.9	13.2	79.7
	DoLa	55.6	17.0	97.1	52.6	15.2	88.8	46.6	13.6	80.3
	CGD (ours)	**42.7**	**10.9**	99.6	**35.7**	**8.6**	85.1	**29.7**	**8.1**	76.7
NoCaps (Near-Domain)	Greedy	55.7	15.4	102.9	39.5	10.1	77.7	46.1	12.2	80.0
	Nucleus	54.1	14.7	98.9	40.9	10.6	78.0	45.4	11.9	79.5
	TopK	53.2	14.1	98.6	41.6	10.9	78.1	46.6	12.1	79.8
	DoLa	54.5	14.0	97.9	41.0	10.8	77.9	44.8	12.1	79.5
	CGD (ours)	**42.6**	**12.3**	100.6	**29.0**	**6.9**	72.8	**33.3**	**7.9**	75.9
NoCaps (Out-of-Domain)	Greedy	51.3	17.3	98.4	36.4	12.4	68.0	40.2	14.2	69.1
	Nucleus	48.8	16.3	92.8	36.8	12.1	68.8	40.7	14.3	69.4
	TopK	50.3	17.0	92.2	37.1	12.3	68.2	39.9	13.5	69.5
	DoLa	48.0	16.3	92.3	35.9	13.2	68.3	43.5	14.4	71.1
	CGD (ours)	**42.0**	**11.6**	94.1	**26.6**	**8.4**	65.7	**28.8**	**9.0**	66.3

open-sourced multi-modality task benchmark MM-Vet (Yu et al., 2023) and follow the provided automatic evaluation procedure. See more details in Appendix C.

Metrics. We follow the guidelines in (Rohrbach et al., 2018) to calculate CHAIR metrics for automatic hallucination evaluation. More precisely, CHAIR quantifies the degree of object hallucination in a given image description by computing the ratio of all objects mentioned in the description but not present in the ground-truth label set. It comprises two assessment dimensions: CHAIR_s (C_S) calculated at the sentence-level, and CHAIR_i (C_I) calculated at the instance-level. Specifically, CHAIR computes CHAIR_i and CHAIR_s as follows:

$$\text{CHAIR}_i = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|}, \quad \text{CHAIR}_s = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}.$$

Following (Rohrbach et al., 2018), we restrict to the objects in 80 COCO object classes for the COCO dataset. For NoCaps, we set a similar setting and map the fine-grained classes defined in NoCaps to coarse-grained categories based on the hierarchical object relationships in Open Images to improve the effectiveness of CHAIR metrics. We only add the super-categories defined in Open Images to our final object list. Eventually, we construct a list of 90 coarse-grained object categories. We include more details in the Appendix C.

4.1 Experimental Results

Hallucination Evaluation on COCO and NoCaps. As COCO has been prevalently used in fine-tuning LVLMs (Liu et al., 2023b; Dai et al.), we extend our evaluation to include images from NoCaps dataset that feature object classes less or not presented in COCO. This set includes near-domain and out-of-domain images. Table 1 displays the empirical results with CHAIR metrics for COCO, near-domain and out-of-domain data. Compared to responses generated in COCO, the average response length of out-of-domain images is shorter, indicating the LVLM models generally output less for their less confident data. As a smaller CHAIR metric reflects a lower fraction of hallucinated objects generated in the response, our method consistently surpasses other baselines, achieving lower scores in the CHAIR metrics while maintaining response lengths similar to those produced by the baselines.

Generation Quality Evaluation on COCO. An effective decoding strategy is crucial for minimizing hallucinations while maintaining high-quality generation in textual outputs. To assess response quality in LVLMs, we expand our evaluation to include widely recognized caption-related metrics such as CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016). We calculate these metrics for responses generated by three different models on the COCO dataset. As shown in Table 2, our

Table 2: Generation quality evaluation on COCO Karpathy Test Split (Karpathy & Fei-Fei, 2015).

Model		Avg. Length	Avg. Coverage	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	SPICE	CLIPS
InstructBLIP	Greedy	102.65	81.10	15.85	10.97	7.02	4.41	17.12	16.99	17.69	27.06
	Nucleus	98.30	80.61	16.38	11.29	7.21	4.53	17.45	17.46	18.17	27.05
	TopK	97.26	80.04	16.46	11.26	7.13	4.43	17.36	17.48	18.00	27.03
	DoLa	97.11	80.11	16.54	11.34	7.22	4.50	17.47	17.61	18.11	27.07
	CGD	99.64	79.44	16.38	11.30	7.19	4.50	17.44	17.49	18.46	28.24
mPLUG-Owl2	Greedy	89.35	81.28	18.08	12.59	8.28	5.39	18.79	19.28	19.18	27.08
	Nucleus	88.98	81.68	18.12	12.60	8.26	5.36	18.78	19.38	19.30	27.05
	TopK	89.15	81.36	18.09	12.51	8.18	5.29	18.76	19.38	19.22	27.06
	DoLa	88.76	81.30	18.14	12.58	8.20	5.29	18.81	19.46	19.20	27.07
	CGD	85.06	80.17	19.07	13.42	8.88	5.77	19.37	20.35	20.21	28.21
LLaVA-1.5	Greedy	80.05	81.30	18.91	12.73	8.07	5.07	18.69	20.23	18.49	26.94
	Nucleus	80.14	81.19	18.88	12.77	8.10	5.12	18.73	20.25	18.50	26.93
	TopK	79.71	81.39	18.90	12.69	8.05	5.07	18.66	20.20	18.39	26.92
	DoLa	80.29	80.88	18.74	12.46	7.83	4.88	18.55	19.98	17.98	26.81
	CGD	76.66	79.03	19.85	13.46	8.60	5.43	19.34	21.13	19.37	27.96

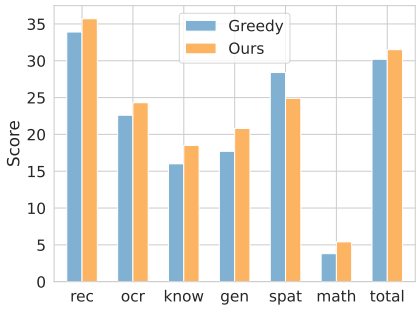


Figure 2: Results with LLaVA-1.5 on MM-Vet (Yu et al., 2023) to evaluate integrated capabilities on multi-modality tasks: recognition (rec), ocr, knowledge (know), language generation (gen), spatial awareness (spat) and math. Please see the case study in Table 7.

method performs on par with other approaches across these diverse metrics. This underscores our method's ability to preserve the overall utility of text generation, while reducing hallucination.

Open-Ended VQA Performance on MM-Vet. In addition to evaluating hallucination, we conduct a study on our algorithm using the open-ended VQA benchmark, MM-Vet (Yu et al., 2023), to gain insights into its effect in a wider array of settings. Figure 2 illustrates the empirical comparison between our method and the standard greedy decoding approach. Notably, our method surpasses the baseline in most tasks, such as recognition and OCR. However, it underperforms in spatial awareness tasks, likely attributable to the limitations in relational understanding of CLIP models (Thrush et al., 2022; Yuksekgonul et al., 2022; Hsieh et al., 2023). Ongoing research seeks to enhance vision-language models through data curation (Fang et al., 2023a; Xu et al., 2023) or feature fusion techniques (Wang et al., 2023b; Tong et al., 2024). These advances could further augment our method through a better guidance model with enhanced visual grounding capabilities.

We include the other evaluation results about ablation study in scoring function and sensitivity study of hyperparameters in Appendix D.

5 Conclusion

In this study, we focus on object hallucination analysis and mitigation for open-ended generation in Large Vision-Language Models (LVLMs). We reveal that there exists severe hallucination in later sentences, and CLIPScore is a stronger and more robust indicator of hallucination than token likelihood. Motivated by this, we integrate CLIP as external guidance in a training-free approach during decoding. Our approach effectively mitigates hallucination while preserving generation quality.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 8948–8957, 2019.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. *Computer Vision–ECCV 2016*, pp. 382–398, 2016.

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Chi Kit Cheung. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. *arXiv preprint arXiv:2204.01171*, 2022.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3), 2023.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*, 2023.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2023.

W Dai, J Li, D Li, AMH Tiong, J Zhao, W Wang, B Li, P Fung, and S Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arxiv 2023. *arXiv preprint arXiv:2305.06500*.

Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. Data filtering networks. *arXiv preprint arXiv:2309.17425*, 2023a.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19358–19369, 2023b.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. *arXiv preprint arXiv:2308.06394*, 2023.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*, 2023.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, et al. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2(3):18, 2017.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 1(2):9, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*, 2023.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437.

Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *arXiv preprint arXiv:2401.06209*, 2024.

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

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. Vigc: Visual instruction generation and correction. *arXiv preprint arXiv:2308.12714*, 2023a.

Haoxiang Wang, Pavan Kumar Anasosalu Vasu, Fartash Faghri, Raviteja Vemulapalli, Mehrdad Farajtabar, Sachin Mehta, Mohammad Rastegari, Oncel Tuzel, and Hadi Pouransari. Sam-clip: Merging vision foundation models towards semantic and spatial understanding. *arXiv preprint arXiv:2310.15308*, 2023b.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023c.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*, 2023.

Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023.

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

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2022.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6720–6731, 2019.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A Smith. How language model hallucinations can snowball. *arXiv preprint arXiv:2305.13534*, 2023.

Yi-Fan Zhang, Weichen Yu, Qingsong Wen, Xue Wang, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Debiasing large visual language models. *arXiv preprint arXiv:2403.05262*, 2024.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13041–13049, 2020.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

A Related Works

A.1 Large Vision-Language Models

Recent developments in Large Vision-Language Models (LVLMs) (Liu et al., 2023b; Li et al., 2023a; Ye et al., 2023) have been significantly powered by the open-sourcing of Large Language Models (LLMs) such as LLaMA (Touvron et al., 2023) and Vicuna (Chiang et al., 2023). These advancements enable LVLMs to make remarkable strides in understanding and addressing a wide range of vision-language tasks with more integrated capabilities (Yu et al., 2023; Yue et al., 2023). Most LVLMs share the same two training phases, i.e., pre-trained feature alignment and instruction fine-tuning, to align the vision feature with language features from LLMs and make the model comprehend and follow the instruction (Liu et al., 2023b; Dai et al.). While these LVLMs have shown promising improvements in handling more complex and general tasks compared to earlier, smaller vision-language models (Zhou et al., 2020), they are still easily suffered from hallucination issues (Li et al., 2023b; Zhou et al., 2023; Wang et al., 2023c), especially in open-ended generation contexts (Zhang et al., 2023).

A.2 Hallucination in VLMs

While the issue of hallucinations in LLMs has been widely studied in the field of NLP (Ji et al., 2023), hallucination mitigations in recent LVLMs are still unexplored. There are recent efforts in mitigating hallucination, including robust instruction tuning (Liu et al., 2023a) and using intrinsic information from models, such as likelihood scores or hidden states from the model (Zhou et al., 2023; Huang et al., 2023). However, only relying on the internal states of models could be potentially unreliable due to the known overconfident issues in neural networks (Kadavath et al., 2022; Xiong et al., 2023). While leveraging external knowledge or models is an alternative way, existing methods (Yin et al., 2023) often result in substantial engineering complexity. On the other hand, while vision-language pre-trained models, e.g. CLIP, are widely adopted in evaluation (Hessel et al., 2021) and mainly studied in the pairwise comparison context (Yuksekgonul et al., 2022; Thrush et al., 2022; Hsieh et al., 2023), the efficacy of CLIP models in detecting hallucination in open-ended generation remains underexplored. Motivated by this, we first conduct hallucination analysis with likelihood-based scores and CLIP scores at the sentence level.

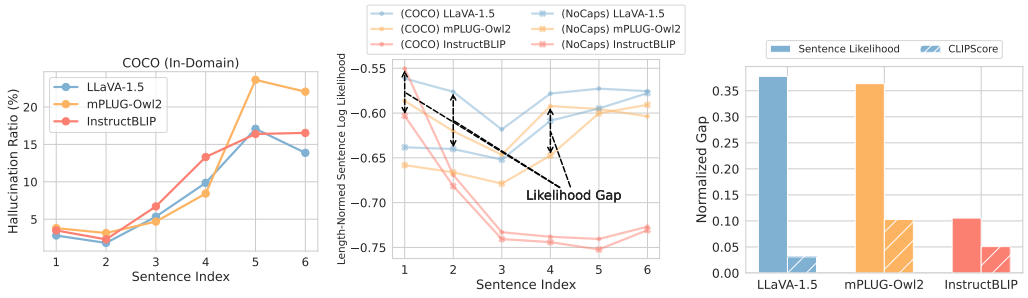
B Hallucination Analysis

In this section, we study how well the sentence likelihood scores and CLIPScore can detect hallucinated sentences generated by LVLMs. Note that recent efforts (Zhou et al., 2023) have also studied uncertainty-related metrics for object hallucination, but focus on the token level, whereas our work focuses on sentence-level hallucination.

As individual tokens lack sufficient semantic meaning, it is challenging to determine if they are hallucinations without broader context. This issue also necessitates careful token selection in practical applications (Zhou et al., 2023). In contrast, a sentence, as a more self-contained unit in natural language, offers a clearer and more effective basis for studying hallucinations.

Models. We have included three LVLMs: InstructBLIP (Dai et al.), mPLUG-Owl2 (Ye et al., 2023) and LLaVA-1.5 (Liu et al., 2023b) in the study. We use Greedy Decoding to generate the response with an image and the prompt, 'Describe this image in detail'. Note that in our work, we use InstructBLIP (Vicuna-7B), mPLUG-Owl2 (LLaMA-7B) and LLaVA-1.5 (Vicuna-7B) and set the maximum new tokens as 500 by default.

Datasets. We use images from COCO (Karpathy & Fei-Fei, 2015) and NoCaps Validation (Out-of-Domain) datasets (Agrawal et al., 2019). Specifically, The NoCaps (Agrawal et al., 2019) dataset is proposed to evaluate models trained on the training set of COCO captions to examine how well they generalize to a much larger variety of visual concepts, i.e., unseen objects. NoCaps (Out-of-Domain) set is a subset of NoCaps dataset and includes samples with novel classes, which are unseen in the COCO dataset. Following the procedure in Rohrbach et al. (2018), the hallucination label $H(\boldsymbol{s}_i)$ for a sentence $\boldsymbol{s}_i$ is decided based on whether the sentence contains a non-existent object label when compared with all objects in the ground-truth caption and segmentation labels provided by



(a) (b) (c)

Figure 3: **(a)** Hallucination ratios in sentences generated by three LVLM models on COCO. The later-generated sentences are more prone to hallucinations. **(b)** Likelihood gap between COCO and NoCaps over three models implies the potential instability of likelihood-only methods when applied to diverse datasets. **(c)** Normalized gap between COCO and NoCaps using sentence likelihood and CLIPScore. A smaller gap value indicates more stable values across datasets. This finding shows that CLIPScore values are more stable across datasets compared to sentence likelihood.

the datasets. We include 1000 samples in analysis for each dataset. For object labels in datasets, we include more details in Appendix C. The object labels are restricted to 80 classes of COCO dataset. For NoCaps, we set a similar setting and map the fine-grained objects defined in NoCaps to coarse-grained categories based on the hierarchical object relationships in Open Images[2] to improve the effectiveness of evaluation. We only add the super-categories defined in Open Images to our final object list. Eventually, we construct a list of 90 coarse-grained object categories. We include more details in the Appendix.

Metrics To quantify hallucination at sentence level, we define two metrics: hallucination ratio $R(\cdot)$ and first-time hallucination ratio $R_{\text{first}}(\cdot)$. For evaluating the detection performance, we use AUROC metric.

Given a dataset $\mathcal{D}$ and the corresponding output $\mathcal{Y}$ generated by a model, we denote $\mathcal{Y}_i := \{\boldsymbol{y} \in \mathcal{Y} \mid |\boldsymbol{y}| \geq i\}$ to be the set of responses where each response $\boldsymbol{y}$ has a length of at least i sentences; here $|\boldsymbol{y}|$ denotes the number of sentences in the response $\boldsymbol{y}$. Hence, we define the *hallucination ratio* at i-th sentence as the fraction of times the i-th sentence is hallucinated, among responses with at least i sentences:

$$R(i) := \frac{|\{\boldsymbol{y} \in \mathcal{Y}_i \mid H(\boldsymbol{s}_i) = 1\}|}{|\mathcal{Y}_i|}. \tag{4}$$

In addition, we define the *first-time hallucination ratio* as the fraction of responses where hallucination first occurs at the i-th sentence, among responses with at least i sentences:

$$R_{\text{first}}(i) := \frac{|\{\boldsymbol{y} \in \mathcal{Y}_i \mid H(\boldsymbol{s}_i) = 1, H(\boldsymbol{s}_j) = 0, \forall j < i|}{|\mathcal{Y}_i|}. \tag{5}$$

Later Sentences Are More Prone to Hallucinations. We investigate the hallucination ratios $R(\cdot)$ across sentences regarding the position of the sentences in the generated description. Figure 3a shows that the sentences generated in the later part are more prone to hallucination, with surprisingly consistent increasing pattern across multiple LVLMs. This finding echoes previous observations about object positional bias in Zhou et al. (2023), which is at the token level. This bias indicates the severity of hallucination as longer descriptions are generated.

Does this positional bias occur solely due to error propagation in sequential generation (Arora et al., 2022; Zhang et al., 2023), i.e., early errors inducing later errors? To investigate this, we focus on first-time hallucination ratios $R_{\text{first}}(\cdot)$ across different sentence indexes, which removes the effect of error propagation. Interestingly, as shown in Figure 6, the bias remains evident. This suggests that positional bias is not exclusively a result of error propagation. Instead, it may be partly attributed to diminishing attention to visual inputs as the length of the generated descriptions increases Zhang et al. (2024).

[2] https://storage.googleapis.com/openimages/web/download_v7.html#df-classes-hierarchy

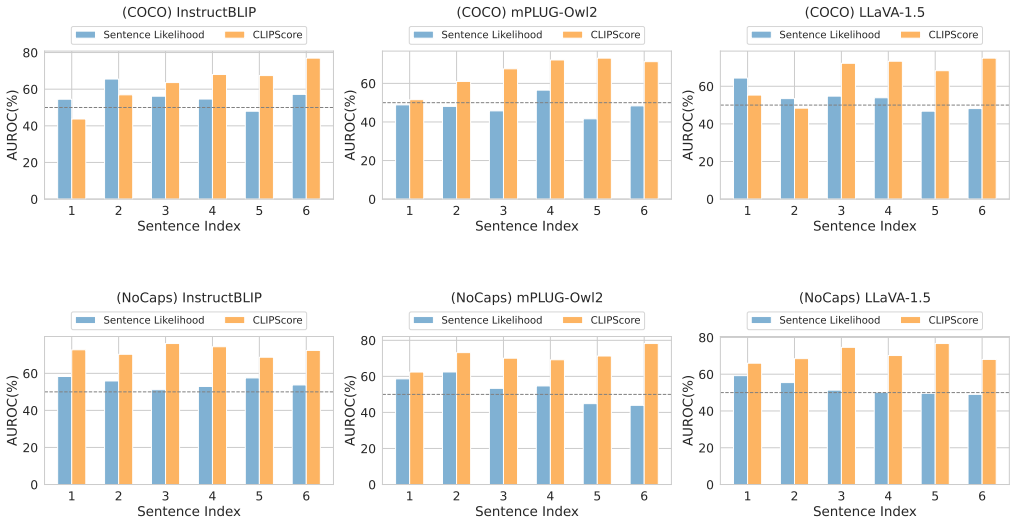
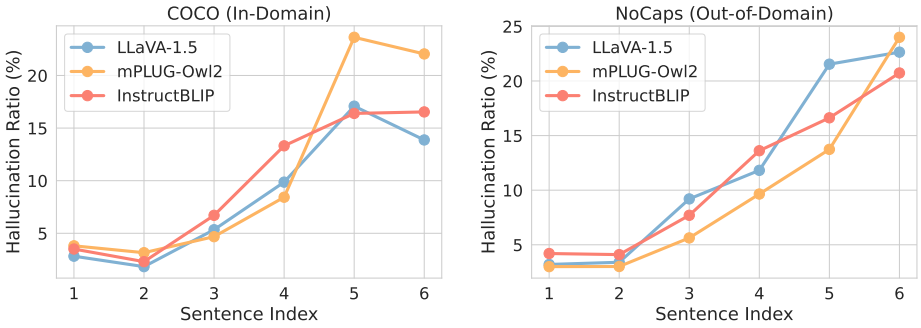


Figure 4: Hallucination detection performance (AUROC) of using Sentence Likelihood $g_{\boldsymbol{\theta}}(\cdot)$ and CLIPScore over COCO and NoCap (out-of-domain) datasets with three LVLMs. CLIPScore outperforms sentence likelihood generally, especially for the later sentences. An AUROC of 50% represents random guessing.

Poor Predictive Performance of Sentence Likelihood. Next, we assess the performance of hallucination detection, measured by AUROC, using length-normalized sentence likelihood $g_{\boldsymbol{\theta}}(\cdot)$ as defined in Eq. (1). An effective metric should distinguish between hallucinated and non-hallucinated sentences, yielding a higher AUROC value. As depicted in Figure 4, the performance of sentence likelihood diminishes in later sentences and exhibits generally weak predictive capabilities in hallucination detection.

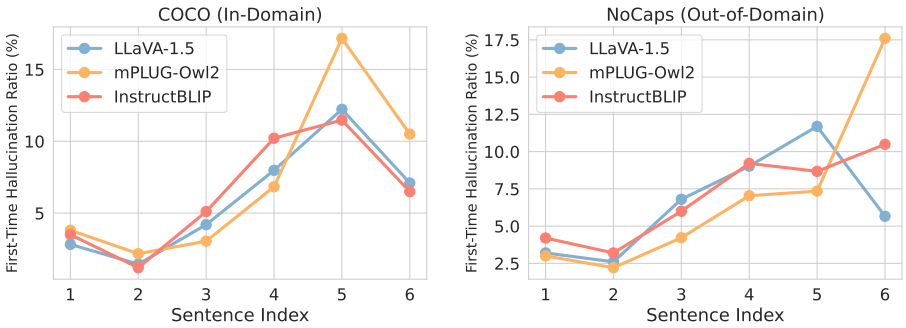
Likelihood Gap of Sentence Likelihood. To investigate the stability of likelihood scores over different datasets, in Figure 3b we compare the mean values of sentence likelihood across COCO and NoCaps (Out-of-Domain) datasets. Notably, a discernible gap often exists between the mean likelihood scores over COCO and out-of-domain data. Specifically, likelihood scores obtained in COCO are consistently higher than those in the out-of-domain data, suggesting that models may exhibit more confidence in the data they were trained on, particularly since COCO datasets are commonly utilized for fine-tuning in LVLMs (Liu et al., 2023b; Dai et al.). This observation raises concerns about relying solely on likelihood scores in real-world applications, as these internal metrics from models may be influenced by training data or other model-specific characteristics (Ranzato et al., 2015), making it challenging to generalize them to out-of-domain data.

Can CLIPScore Detect Hallucination? Existing studies predominantly focus on CLIPScore within a pairwise context, wherein a correct caption is compared with a modified version generated through word manipulation (such as removal, addition, or swapping) (Thrush et al., 2022; Hsieh et al., 2023). There is limited exploration of CLIPScore's efficacy in detecting hallucinations within an open-world generation setting. In contrast to the traditional pairwise approach, our investigation demonstrates that CLIPScore is adept at distinguishing incorrect (hallucinated) sentences from correct ones in a broader context. Illustrated in Figure 4, CLIPScore exhibits notable effectiveness in distinguishing hallucinated and non-hallucinated sentences across different models and datasets. Moreover, as CLIP models function as independent external examiners, they exhibit insensitivity to positional bias, performing well across sentence indexes. We also investigate the stability of scores across different datasets, specifically COCO and NoCaps. Figure 3c indicates that CLIPScore maintains greater consistency across these datasets when compared to sentence likelihood scores.



(a) (b)

Figure 5: Hallucination ratios across different sentence indexes on both COCO and NoCaps (Out-of-Domain) datasets.



(a) (b)

Figure 6: First-time hallucination ratios across different sentence indexes on both COCO and NoCaps (Out-of-Domain) datasets.

C EXPERIMENTS

C.1 MODELS

We select three representative LVLM models for evaluation, including InstructBLIP (Dai et al.), mPLUG-Owl2 (Ye et al., 2023) and LLaVA-1.5 (Liu et al., 2023b). InstructBLIP adopts a Q-former to bridge the features between the vision and text modalities, using 32 tokens as image token embeddings. mPLUG-Owl2 and LLaVA-1.5 use linear projection layers to align the vision and text modalities, with 256 or 576 tokens as image representations. Generally, the image encoders of these LVLM models are pre-trained models trained with image-text contrastive objective, e.g., CLIP (Radford et al., 2021) and EVA-CLIP (Fang et al., 2023b). Note that we use InstructBLIP (Vicuna-7B), mPLUG-Owl2 (LLaMA-7B) and LLaVA-1.5 (Vicuna-7B) and set the maximum new tokens as 500 by default for all methods.

C.2 DATASETS

COCO. The COCO dataset (Lin et al., 2014) is a comprehensive dataset used for image recognition, segmentation and captioning. It contains over 300,000 images spanning over 80 object categories, each with detailed annotations. Given the detailed and high-quality annotation, most recent LVLMs use samples from COCO for vision-language alignment and instruction tuning (Liu et al., 2023b; Li et al., 2023a). Specifically, we use samples from COCO Karpathy Split (Karpathy & Fei-Fei, 2015) in our experiments.

Algorithm 1 CLIP-Guided Decoding

Input: input containing an image and textual prompt $\boldsymbol{x} := (\boldsymbol{x}_{\text{img}}, \boldsymbol{x}_{\text{text}})$, LVLM parameterized by $\boldsymbol{\theta}$, CLIP model parameterized by $\boldsymbol{\phi}$, maximum cardinality N, sampling size M, weight hyperparameter α
Output: Response output $\mathbf{c}^*$
$t := 0$
$\mathcal{C}_0 := \{\langle\text{start token}\rangle\}$.
while not EOF **do**
  $\mathcal{C}'_{t+1} := \varnothing$
  for $\mathbf{c}_i^t \in \mathcal{C}_t$ **do**
    repeat
      $\boldsymbol{s} \sim \mathbb{P}_\theta(\boldsymbol{s}_{t+1} \mid \boldsymbol{x}, \boldsymbol{s}_{i,1}^t, \ldots, \boldsymbol{s}_{i,t}^t)$
      $\mathcal{C}'_{t+1} := \mathcal{C}'_{t+1} \bigcup \{(\boldsymbol{s}_{i,1}^t, \ldots, \boldsymbol{s}_{i,t}^t, \boldsymbol{s})\}$
    until M times
  end for
  $\mathcal{C}_{t+1} :=$ Top N candidates in $\mathcal{C}'_{t+1}$
  $t := t + 1$
end while
$\mathbf{c}^* := \arg\max_{\mathbf{c} \in \mathcal{C}_t} F(\mathbf{c})$ ▷ Eq. (3)
Return $\mathbf{c}^*$

NoCaps. The NoCaps dataset (Agrawal et al., 2019) is proposed for evaluating the models trained with COCO with images less or not seen in the COCO object categories. There are $4,500$ images in the validation set and $10,600$ images in the test set. Images are taken from the Open Images V4 (Krasin et al., 2017) dataset, which spans 600 object classes. Due to the unavailability of ground truth captions of the test set, we use the validation set of NoCaps.

MM-Vet. MM-Vet (Yu et al., 2023) is an evaluation benchmark[3] that examines large multimodal models on complicated multimodal tasks. It defines 6 core vision-language capabilities including recognition, OCR, knowledge, language generation, spatial awareness and math. It examines the 16 integrations of interest derived from the combination of these 6 capabilities. In total, this benchmark contains 200 images and 218 questions, all paired with their respective ground truths, which are human annotated or gathered from the internet. Specifically, the benchmark has gathered 187 images from various online sources, 10 high-quality images from VCR (Zellers et al., 2019) and 3 images from ChestX-ray14 (Wang et al., 2017). The benchmark has also labeled the capacities required for answering each question correctly.

C.3 METRICS.

Hallucination Evaluation. We follow the guidelines in (Zhou et al., 2023) to calculate CHAIR metrics for automatic hallucination evaluation. More precisely, CHAIR quantifies the degree of object hallucination in a given image description by computing the ratio of all objects mentioned in the description but not present in the ground-truth label set. It comprises two assessment dimensions: CHAIR_s (C_S) calculated at the sentence-level, and CHAIR_i (C_I) calculated at the instance-level. Specifically, CHAIR computes CHAIR_i and CHAIR_s as follows:

$$\text{CHAIR}_i = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|},$$
$$\text{CHAIR}_s = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}.$$

Following Rohrbach et al. (2018); Zhou et al. (2023), we restrict the objects in 80 COCO object classes for the COCO dataset. For NoCaps, we set a similar setting and map the fine-grained classes defined in NoCaps to coarse-grained categories based on the hierarchical object relationships in Open Images[4] to improve the effectiveness of CHAIR metrics. Specifically, we only add the super-

[3] https://github.com/yuweihao/MM-Vet

[4] https://storage.googleapis.com/openimages/web/download_v7.html#df-classes-hierarchy

categories defined in Open Images to our final object list. Eventually, we construct a list of 90 coarse-grained object categories from 600 fine-grained object classes.

Generation Quality Evaluation. To reflect the quality of the generated description, we include several metrics: average word number in a response (Avg. Length) and average coverage ratio (Avg. Coverage), which calculates the proportion of correctly identified objects in a generated description relative to the total number of 'golden' (i.e., actual or reference) objects present in the image. We also include other caption-related evaluation metrics including BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015), SPICE (Anderson et al., 2016) and CLIPScore (Hessel et al., 2021):

- **BLUE** BLUE (Bilingual Evaluation Understudy (Papineni et al., 2002)) is a metric for evaluating the quality of machine-generated translations by comparing them to one or more reference translations. The BLEU score is based on precision of n-grams, which are contiguous sequences of n words.
- **METEOR** METEOR (Metric for Evaluation of Translation with Explicit ORdering (Banerjee & Lavie, 2005)) is designed to evaluate the quality of machine-generated text (translations or captions) by comparing it to one or more human reference texts. Specifically, METEOR computes a harmonic mean (F-mean) of precision and recall, incorporating both the quality of the generated text and its similarity to reference text.
- **ROUGE-L** ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence (Lin, 2004)) is a common metric in evluating text summarization tasks. It is proposed to measure the quality of a machine-generated summary by comparing it to one or more reference summaries.
- **CIDEr** CIDEr (Consensus-based Image Description Evaluation (Vedantam et al., 2015)) is designed for image captioning evaluation by measuring the quality of generated image captions by comparing them to human-generated reference captions. The score utilizes the weighted combination of n-gram similarity scores in both the generated and reference captions. Specfically, it gives more weight to higher-order n-grams to encourage diversity in generated captions.
- **SPICE** SPICE (Semantic Propositional Image Caption Evaluation (Anderson et al., 2016)) metric is an evaluation measure used for image captioning. Instead of focusing on n-grams or surface-level text similarity, SPICE parses both the generated and reference captions into semantic propositions, including objects, attributes, relationships, and actions present in the image. It aims to capture the accuracy of generated captions with an aspect of semantic meanings.
- **CLIPScore** CLIPScore (Radford et al., 2021; Hessel et al., 2021) is a metric based on CLIP (Contrastive Language-Image Pretraining (Radford et al., 2021)) to measure how well a given textual description is associated with the image by computing the cosine similarity based on the normalized features through CLIP models.

Open-Ended VQA Benchmark Evaluation on MM-Vet. For evaluation of open-ended generation, due to the high flexibility and free-form of answers, MM-Vet has proposed an LLM-based evaluator for open-ended outputs. Specifically, they provide a template with the question, ground-truth answer and prediction (e.g., output from an LVLM) in a few-shot prompt and prompt the LLMs to provide a soft grading score from 0 to 1, where a high score indicates a more accurate prediction. The total scores are computed by the average scores obtained for all questions. The score regarding each capability is the average score obtained for the questions that have been annotated for the requirement of the specific capability.

D EXPERIMENTAL RESULTS

Ablation Study of Hallucination Scoring. To further understand the functionality of sentence likelihood and CLIPScore in our hallucination scoring, defined in Eq. (3), we conduct an ablation study focusing on sentence likelihood and CLIPScore's impact on hallucination mitigation by controlling α. Specifically, the algorithm operates without sentence likelihood when $\alpha = 1$ and without CLIP

Table 3: Ablation study on scoring components in Eq. (3).

Method	$C_S \downarrow$	$C_I \downarrow$	Avg. Len
Greedy	44.67	13.11	80.05
CGD (Ours)	29.73	8.12	76.66
w/o Sentence Likelihood	**29.45**	**7.37**	76.12
w/o CLIP-Guidance	45.07	13.29	81.60

guidance when $\alpha = 0$. From Table 3, we can see that the performance without CLIP-guidance is close or even slightly worse than greedy decoding and sentence likelihood demonstrates relatively limited effectiveness in reducing hallucination. This result confirms the dominant contribution of CLIP guidance in hallucination mitigation.

Table 4: Sensitivity analysis of maximum candidate number N and sampling times M.

Method	N	M	$C_S \downarrow$	$C_I \downarrow$	Avg. Len
Greedy	-	-	44.67	13.11	80.05
CGD	1	3	37.66	9.90	78.67
	1	5	34.50	8.95	78.70
	3	3	29.73	8.12	76.66
	5	3	**28.40**	**6.78**	78.18

Sensitivity Analysis of Hyperparameters N and M. To examine the sensitivity of our method to variations in maximum candidate number N and sampling times M, we carry out the experiments with different settings where $N \in \{1, 3, 5\}$ and $M \in \{3, 5\}$. We omit $M = 1$ from our study as this setting reverts to the ordinary sampling. Table 4 shows that our method outperforms the baseline even with a reduced maximum candidate number and sampling times. Generally, the performance improves with larger N and M. This suggests that having a broader range of choices and larger candidate budgets helps the model to avoid hallucination.

Table 5: Sensitivity analysis of CLIP-guidance model.

Method	$C_S \downarrow$	$C_I \downarrow$	Avg. Len
Greedy	44.67	13.11	80.05
w/ SigLIP ViT-SO-14@384px	**29.73**	**8.12**	76.66
w/ OpenAI ViT-L-14@336px	32.80	8.78	75.83

Effect of CLIP Guiding Model. As most of the LVLMs have adopted CLIP models from OpenAI as vision encoders, it is worthwhile to investigate whether utilizing the same CLIP models as guiding models in our method can still enhance hallucination mitigation. For a fair comparison, we maintain all the settings but substitute the guidance model with CLIP ViT/L-14 with 336px resolution image input (Radford et al., 2021), which is used as vision encoder in LLaVA-1.5. Table 5 demonstrates that our method continues to surpass the baseline method even when using this vision encoder. This finding suggests that the existing fine-tuning processes for vision-language alignment in LVLMs (Liu et al., 2023b; Dai et al.) might, to some extent, compromise the original vision capabilities. Our method is a way to recalibrate the alignment between vision and language by directly examining the outputs from LVLMs.

D.1 Case Study

Hallucination Case Study. We have included several cases of hallucination mitigation on the COCO dataset, as shown in Table 6. This table shows examples by comparing the original generated description and generated description by our method given the image in the leftmost column. The hallucinated part in the description has been highlighted in red.

Table 6: Examples of COCO with LLaVA-1.5. Comparison between the text generated by Original (TopK Sampling with $K = 5$) and our methods. The hallucinated parts are highlighted in red.

Image	Original	Ours
	The image depicts a man sitting at a dining table, enjoying a meal of pizza. He is wearing a police uniform, indicating his occupation. There are two pizzas on the table, one closer to the man and the other further away. The table is set with a cup, a fork, and a knife. A book is also present on the table, possibly for the man to read while he eats. The scene is set in a cozy environment, with a couch visible in the background.	The image features a man wearing a police uniform sitting at a dining table with a plate of pizza in front of him. The table is set with two cups, one on the left side and the other on the right side. There are also two slices of pizza on the table, one in front of the man and the other on the right side. The man appears to be enjoying his meal, and the scene seems to be a casual dining experience.
	The image features a small dog standing on a paved road, wearing a collar. The dog is positioned near the center of the scene. A motorcycle is parked on the right side of the dog, and a bus is visible in the background, occupying a large portion of the image. There are two people in the scene. One person is standing near the left edge of the image, while the other person is located on the right side, closer to the motorcycle.	The image features a small brown and white dog standing on a paved surface, possibly a parking lot. The dog is on a leash, and it appears to be looking at the camera. In the background, there is a bus parked, and a motorcycle is also visible in the scene. The dog is positioned near the center of the image, with the bus and motorcycle occupying the right side of the frame.
	The image features a woman wearing a blonde wig and holding a tennis racket. She is smiling and appears to be enjoying herself. The woman is standing in a room with a couch in the background. There is a TV mounted on the wall, and a cup can be seen placed on a surface nearby. The overall atmosphere of the scene is light-hearted and fun.	The image features a woman wearing a blonde wig and a green shirt, holding a tennis racket. She appears to be dressed as a tennis player, possibly for a costume party or a themed event. The woman is smiling and posing with the tennis racket, showcasing her enthusiasm for the sport.

VQA Case Study. We have demonstrated some examples from the MM-Vet dataset in Table 7. We have inlcuded some samples with each row containing the image, the corresponding question, the originally generated response, a response based on our method and the provided ground-truth answer. The score under each response reflects the accuracy of the response to the question. We also show the required capabilities for answering each question in the "Answer" Column.

Table 7: Examples from MM-Vet (Yu et al., 2023) based on LLaVA-1.5.

Image	Question	Original	Ours	Answer
	Can you give a short introduction to this painting?	This is a painting of a night scene with a city in the foreground and a mountain in the background. The painting is a representation of the night sky with the stars and the moon. The city is depicted with buildings and the mountain is shown with a church on top. The painting is a beautiful representation of the night sky and the cityscape. **(score: 0.2)**	This painting is a famous representation of the night sky by the artist Vincent Van Gogh. The painting is a depiction of the night sky with the stars visible, and it is set against a backdrop of a town. The painting is a combination of the artist's imagination and the night sky, creating a unique and captivating piece of art. **(score: 0.6)**	The Starry Night is an oil-on-canvas painting by the Dutch Post-Impressionist painter Vincent van Gogh. Painted in June 1889, it depicts the view from the east-facing window of his asylum room at Saint-Rémy-de-Provence, just before sunrise, with the addition of an imaginary village. It has been in the permanent collection of the Museum of Modern Art in New York City since 1941, acquired through the Lillie P. Bliss Bequest. Widely regarded as Van Gogh's magnum opus, The Starry Night is one of the most recognizable paintings in Western art. **Required capabilities: recognition, knowledge, language generation**
	Which animal in the picture is larger, the cat or the dog?	The cat is larger than the dog. **(score: 1.0)**	The dog is larger than the cat in the picture. **(score: 0.0)**	cat **Required capabilities: recognition, spatial awareness**
	What should kids do after snap fingers?	After snap fingers, kids should then touch their toes. **(score: 0.0)**	After snap fingers, kids should hop on one foot. **(score: 1.0)**	hop on one foot **Required capabilities: OCR, spatial awareness**
	What is the name of this dish?	The name of this dish is "Tofu Soup" or "Tofu Stew." **(score: 0.0)**	The dish is called "Tofu in a spicy sauce". **(score: 0.4)**	mapo tofu **Required capabilities: recognition**