ECD: Efficient Contrastive Decoding with Probabilistic Hallucination Detection

Laura Fieback✉[1,2][0009-0003-3456-6766], Nishilkumar Balar[1][0009-0008-7815-9536], Jakob Spiegelberg[1][0000-0002-6550-0087], and Hanno Gottschalk[2][0000-0003-2167-2028]

[1] Volkswagen AG, Berliner Ring 2, 38440 Wolfsburg, Germany {laura.fieback, nishilkumar.balar, jakob.spiegelberg}@volkswagen.de
[2] TU Berlin, Straße des 17. Juni 136, 10623 Berlin, Germany
gottschalk@math.tu-berlin.de

Abstract. Despite recent advances in Large Vision Language Models (LVLMs), these models still suffer from generating hallucinatory responses that do not align with the visual input provided. To mitigate such hallucinations, we introduce Efficient Contrastive Decoding (ECD), a simple method that leverages probabilistic hallucination detection to shift the output distribution towards contextually accurate answers at inference time. By contrasting token probabilities and hallucination scores, ECD subtracts hallucinated concepts from the original distribution, effectively suppressing hallucinations. Notably, our proposed method can be applied to any open-source LVLM and does not require additional LVLM training. We evaluate our method on several benchmark datasets and across different LVLMs. Our experiments show that ECD effectively mitigates hallucinations, outperforming state-of-the-art methods with respect to performance on LVLM benchmarks and computation time.

Keywords: Multimodal large language models · Contrastive decoding · Hallucination mitigation · Hallucination detection · Meta classification.

1 Introduction

By aligning textual and visual features, LVLMs have shown an impressive vision-language understanding across various multimodal tasks like visual question answering (VQA) or image captioning [31,5,8]. However, inconsistencies between the generated response and the visual input, a phenomenon called hallucinations [35,28], diminish the applicability of LVLMs in safety-critical applications such as autonomous driving [13,45] or medicine [18,26]. Motivated by recent findings [20,15] that identified overreliance of LVLMs on language priors as one of the main reasons for hallucinations, new hallucinatory datasets and fine-tuning strategies have been proposed to mitigate hallucinations [12,11,17]. Contrastive Decoding (CD) strategies [25,43] emerged as a training-free alternative, addressing concerns about computational costs and human effort required for data labeling. The idea of CD is to intervene in the decoding process of LVLMs by amplifying the language prior through distorted inputs and contrasting the output distribution with the distribution derived from original inputs. While this approach effectively mitigates hallucinations and computational overhead, it still increases the inference time by calculating two output distributions.

In this work, we investigate the potential of probabilistic hallucination detection for Efficient Contrastive Decoding (ECD). During the decoding process, token scores are contrasted with hallucination scores to suppress hallucinations. We employ the idea of meta classification [6,24] to train a lightweight detector to estimate hallucination scores based on hallucination features derived from the model output. By investigating features from intermediate LVLM layers, we achieve area under precision recall curve values [7] of up to 74.05%. In contrast to existing CD methods, our approach requires only one forward pass of the LVLM followed by the lightweight hallucination detection, effectively reducing the inference time. Moreover, instead of amplifying hallucinations through input uncertainty, we directly learn hallucinated concepts from internal LVLM calculations, outperforming recent CD methods across several state-of-the-art LVLMs and benchmarks. In detail, ECD mitigates the hallucination rate by up to 5.74*pp*, i.e., 32% in open-ended tasks and improves F1 Scores by 23.02*pp*, i.e., 33% in discriminative VQA benchmarks, while adding only minor computational overhead to the decoding process. Our main contributions are as follows:

- We propose new hallucination features to train a powerful lightweight hallucination detector.
- Based on this detector, we introduce ECD, a lightweight and training-free decoding method that effectively mitigates hallucinations in LVLMs by penalizing mendacious tokens through hallucination scores.
- Through extensive experiments, we show the effectiveness of our approach outperforming state-of-the-art methods on various benchmarks and in computational time.

2 Related Work

2.1 Hallucination Mitigation for LVLMs

The research area of vision-language pre-trained models has made substantial progress by incorporating Large Language Models (LLMs) building the powerful Large Vision Language Models (LVLMs). In general, LVLMs consist of (i) a vision encoder to extract vision features from the input image, (ii) a cross-modal alignment module, which aligns the visual and language features, and (iii) an LLM, which generates the text response. Despite remarkable zero-shot capabilities in multimodal tasks, LVLMs suffer from hallucinations, i.e., they generate answers that do not align with the input image. Several hallucination mitigation methods have been proposed comprising new instruction tuning datasets for LVLM retraining [11,12,17], leveraging expert models for post hoc hallucination correction [46,41], or incorporating object grounding features [23,4,30]. However, these methods require extensive data collection and annotation, LVLM retraining or architecture changes, which can be time-consuming and computationally costly. To cope with this problem, simple contrastive decoding strategies have been introduced, which contrast the output distributions with original and distorted inputs during inference. Based on the observation that hallucinations often occur due to the overreliance of LVLMs on language priors [35], the authors of Visual Contrastive Decoding (VCD) [25] propose to contrast the original output distribution with the distribution derived from noisy input images to subtract the language bias from the original distribution. Similarly, Instruction Contrastive Decoding (ICD) [43] adds prefixes to the text input to increase multimodal alignment uncertainty and finally contrasts the resulting distribution with the original output. Although these methods successfully mitigate hallucinations, they increase the inference time by performing one forward pass with original inputs and one with distorted inputs. Instead, we propose to contrast the output distribution with hallucination scores derived from internal LVLM calculations using meta classification, which effectively reduces computational costs during contrastive decoding.

2.2 Meta Classification for Hallucination Detection

In order to judge the reliability of LVLM responses, different hallucination detection methods have been introduced. These methods either apply a pipeline of stacked LLMs and LVLMs [19,44] to detect hallucinations as a post hoc method or train L(V)LM-based classifiers [20,12] using hand-crafted hallucination datasets. Since these methods are computationally costly, the authors of MetaToken [24] introduced a lightweight and simple hallucination detection method based on meta classification [6]. In general, meta classification refers to the classification of true and false predictions based on uncertainty features derived from the model output. This idea has been applied to various fields like image classification [3], semantic segmentation [32,36,9], video instance segmentation [33], and object detection [22,38]. In [24] new input features for the hallucination detection problem have been proposed that outperform classical uncertainty-based features and can be derived from internal LVLM calculations.

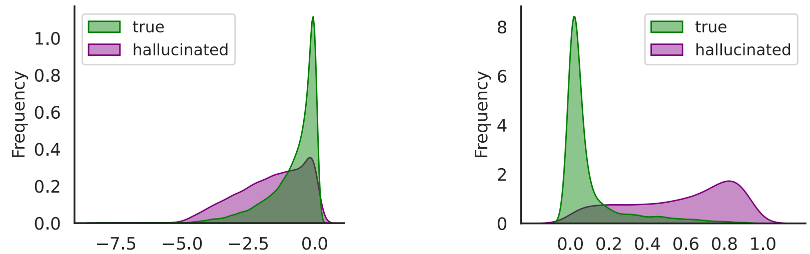
3 Method

3.1 LVLM Decoding

In general, LVLMs generate text responses in an autoregressive way by predicting the probability distribution over the dictionary $\mathcal{V}$ based on the input image v, the input query q, and the sequence already generated. In generation step t, the next token $y_t \in \mathcal{V}$ is generated by sampling from this distribution. Mathematically, this process can be formulated as

$$y_t \sim p_\theta(y_t | v, q, y_{<t}), \tag{1}$$

where θ denote the LVLM parameters and $y_{<t} = (y_0, \ldots, y_{t-1})$ the generated sequence up to generation step $t-1$. Note that a perfect model should assign



(a) log probability $\log p_\theta$ (b) hallucination score p_f

Fig. 1: Visualization of (a) log probability values and (b) hallucination scores for true and hallucinated tokens according to the MSCOCO CHAIR evaluation (see Sec. 4.2).

high probabilities to true tokens and low probabilities to hallucinations. During this decoding mechanism, hallucinations might be generated when tokens with low probabilities are sampled from $p_\theta(y_t|v, q, y_{<t})$. However, as we can see in Fig. 1a, the phenomenon of hallucinations often occurs, as the model assigns high probability values to hallucinated tokens. Our approach corrects this undesired behavior by shifting the final distribution towards true tokens, reducing the probability assigned to hallucinations.

3.2 Probabilistic Hallucination Detection

As we have seen in the previous section, the probability distribution calculated during LVLM decoding does not properly distinguish between true and hallucinated tokens. The idea of meta classification is to learn this classification from hallucination features, which are derived from internal LVLM calculations. By learning from interactions and coherences of these features, the classifier can successfully distinguish between true tokens and hallucinations (see Fig. 1b).

To train our classifier, we build on the work of [24] and extend the set of input features to enhance the hallucination detection capabilities. While the features from [24] are based on the last LLM layer, we integrate further information from the preceding layers into our classifier. This idea is motivated by findings from the LLM literature [47,1] indicating that the middle layers contain information about the reliability of the generated response. To this end, let N denote the number of LLM transformer layers, let $v_0, \ldots, v_u$ denote the image tokens derived from the vision encoder and alignment module, and $q_0, \ldots, q_{w+t}$ the textual tokens representing the input query q and the sequence $y_{<t}$. The concatenated sequence of visual and textual tokens is fed into the LLM and successively processed by each layer i calculating the hidden states $\{h_0^i, \ldots, h_{u+(w+1)+t}^i\}$ with $i = 1, \ldots, N$. Finally, the vocabulary head $\phi(\cdot)$ predicts the probability distribution for the

next token as

$$p_\theta(y_t|v,q,y_{<t}) = \text{softmax}\big[\phi(h^N_{u+(w+1)+t})\big]_{y_t}, \quad y_t \in \mathcal{V}. \tag{2}$$

In order to extract information from the preceding layers, the early exit method [42,34,39] applies the language head to the hidden states of the earlier layers:

$$p^i_\theta(y_t|v,q,y_{<t}) = \text{softmax}\big[\phi(h^i_{u+(w+1)+t})\big]_{y_t}, \quad i \in \{1,\dots,N\}. \tag{3}$$

For a shorter notation, we write p^i_θ. Moreover, let $\text{Att}^{i,g}_{y_t}(j)$ denote the attention on token j in generation step t for layer i and attention head g with $g = 1,\dots,G$. With this notation, we introduce new features based on intermediate layers:

- the negative log-likelihood for all layers

$$B^i(y_t) = -\log p^i_\theta, \quad i = 1,\dots,N \tag{4}$$

- the Kullback–Leibler (KL) [37] divergence between the preceding layers and the last layer

$$K^i(y_t) = \text{KL}(p^N_\theta||p^i_\theta) \coloneqq p^N_\theta \log \frac{p^N_\theta}{p^i_\theta}, \quad i = 1,\dots,N-1 \tag{5}$$

- for each attention head, the entropy of the image attention over the layers

$$E^{\text{layer}}_{v_k,g}(y_t) = -\frac{1}{N}\sum_{i=1}^{N} \text{Att}^{i,g}_{y_t}(v_k)\log \text{Att}^{i,g}_{y_t}(v_k), \quad g = 1,\dots,G \tag{6}$$

  averaged over the image tokens with

$$E^{\text{layer}}_{g}(y_t) = \frac{1}{u+1}\sum_{k=0}^{u} E^{\text{layer}}_{v_k,g}(y_t) \tag{7}$$

- for each layer, the entropy of the image attention over the attention heads

$$E^{\text{head}}_{v_k,i}(y_t) = -\frac{1}{G}\sum_{g=1}^{G} \text{Att}^{i,g}_{y_t}(v_k)\log \text{Att}^{i,g}_{y_t}(v_k), \quad i = 1,\dots,N \tag{8}$$

  averaged over the image tokens with

$$E^{\text{head}}_{i}(y_t) = \frac{1}{u+1}\sum_{k=0}^{u} E^{\text{head}}_{v_k,i}(y_t). \tag{9}$$

We aggregate the features from [24] (see supplementary material A) and our proposed inputs to train the classifier. Let $\mathcal{M}$ denote the set of hallucination features and m_{y_l} the corresponding vector for a generated token $y_l \in \mathcal{V}$. The meta classifier can be defined as

$$f : \mathbb{R}^{|\mathcal{M}|} \to \{0,1\}. \tag{10}$$

Following [24], we use the CHAIR evaluation [35] to extract true ($z_l = 0$) and hallucinated ($z_l = 1$) tokens from LVLM responses to build our training and validation data with standardized inputs m_{y_l} and corresponding labels z_l as

$$\{(m_{y_l}, z_l) \quad | \quad l = 1, \ldots, L\}. \tag{11}$$

Once the classifier is trained, we can detect hallucinations during the LVLM generation process by computing the proposed features and applying the classifier afterwards as

$$f(m_{y_t}) = \begin{cases} 1, & p_f(y_t|v, q, y_{<t}) \geq \tau \\ 0, & p_f(y_t|v, q, y_{<t}) < \tau \end{cases} \tag{12}$$

with the estimated probability $p_f(y_t|v, q, y_{<t})$ for tokens to be hallucinated, referred to as hallucination scores, and the threshold τ controlling the precision-recall ratio. Note that the input vector m_{y_t} can be calculated in an automated manner based on internal LVLM calculations only, without any knowledge of the ground truth data.

3.3 Efficient Contrastive Decoding

By directly learning hallucinated concepts, we suppress the generation of hallucinations during the decoding process without an additional LVLM forward pass. In contrast to existing methods, which model the language bias of LVLMs by generating a second output distribution with distorted inputs, we apply a lightweight classifier on the LVLM output to obtain hallucination scores $p_f(y_t|v, q, y_{<t})$, adding only minor computational overhead to the decoding process. At generation step t, the contrastive distribution is computed by subtracting the hallucination scores from the log probabilities $\log p_\theta(y_t|v, q, y_{<t})$ to penalize hallucinations while maintaining high probabilities for true tokens:

$$\begin{aligned} p_{ecd}(y_t|v, q, y_{<t}) = \text{softmax}\big[&(1+\alpha) \log p_\theta(y_t|v, q, y_{<t}) \\ &- \alpha \log p_f(y_t|v, q, y_{<t})\big], \end{aligned} \tag{13}$$

where α controls the magnitude of hallucination correction. Note that for $\alpha = 0$, p_{ecd} is equal to the initial LVLM distribution. Moreover, our proposed efficient contrastive decoding can be integrated into various decoding strategies such as the standard greedy search, beam search [10], and nucleus sampling [14].

3.4 Adaptive Plausibility Constraint

We follow the implementation of VCD [25] and ICD [43] and incorporate an adaptive plausibility constraint (APC) [27] based on the confidence level of the LVLM distribution to maintain high probabilities for semantically trivial tokens. By refining the final contrastive distribution, APC effectively prevents the generation of implausible tokens, and thus preserves the semantic accuracy of the response. This leads to the final formulation of our proposed decoding strategy:

$$\begin{aligned} y_t \sim p_{ecd}(y_t|v, q, y_{<t}), \quad \text{subject to} \\ y_t \in \mathcal{V}_{\text{head}} = \{y_t \in \mathcal{V} \quad | \quad p_\theta(y_t|v, q, y_{<t}) \geq \beta \max_\omega p_\theta(\omega|v, q, y_{<t})\} \end{aligned} \tag{14}$$

with truncation parameter $\beta \in [0, 1]$, where $\beta = 1$ implements the standard greedy search algorithm.

4 Experimental Setup

4.1 Hallucination Detection

We evaluate the information content of our proposed input features ($|\mathcal{M}| = 169$) to learn the differentiation between true and hallucinated answers. First, we sample $5,000$ images from the MSCOCO [29] validation set and apply the prompt

"Describe all objects in the image."

to generate training and validation data for the probabilistic classifier. This results in approximately 30.000 data points depending on the number of objects generated by the respective LVLM (see [24]). As in [24], we employ a logistic regression (LR) and gradient boosting (GB) classifier, which have shown superior performance compared to small neural networks in previous studies [33], and use the features from [24] as our baseline. In detail, we use the LR[3] classifier with saga solver and the GB[4] classifier both with max_iter = 1000 and scikit-learn version 1.5.2. The detection results are evaluated in terms of accuracy (ACC), area under receiver operator characteristic curve (AUROC) and area under precision recall curve (AUPRC) [7]. We average our results over ten randomly sampled training-validation splits using a ratio of 80% training data and 20% validation data.

4.2 Datasets and Evaluation Metrics

CHAIR: The Caption Hallucination Assessment with Image Relevance (CHAIR) [35] metric is widely used in open-ended image captioning tasks and measures the hallucination and coverage rate of LVLMs by checking extracted objects from the generated response against MSCOCO ground-truth labels. CHAIR is defined on the instance level CHAIR_i and sentence level CHAIR_s as

$$\text{CHAIR}_i = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|}, \text{Coverage} = \frac{|\{\text{mentioned objects}\}|}{|\{\text{labeled objects}\}|}, \tag{15}$$

$$\text{and} \quad \text{CHAIR}_s = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}. \tag{16}$$

For the evaluation of our proposed contrastive decoding method, we sample additional 500 images from the MSCOCO validation set, which do not overlap with the hallucination detection training and validation data.

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

[4] `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingClassifier.html`

AMBER: An LLM-free Multi-dimensional Benchmark (AMBER) [21]. Since our probabilistic classifier was trained on the MSCOCO dataset, which might lead to biased results in the preceding evaluation, we additionally evaluate our method on the AMBER dataset, which covers a more diverse range of object categories. In detail, AMBER covers 337 objects compared to 80 categories for the MSCOCO dataset. The open-ended image captions are again evaluated using CHAIR_i, CHAIR_s, and Coverage metrics.

POPE: The Polling-based Object Probing Evaluation (POPE) [28] is a discriminative VQA benchmark to assess the quality of LVLMs with respect to object hallucinations. In detail, POPE uses the template

"Is there a {object} in the image?"

and applies three different sampling strategies to generate negative prompts, which refer to non-existent objects. The *random* (rand.) sampling chooses the probing objects randomly, *popular* (pop.) samples from high-frequency objects and *adversarial* (adv.) samples among objects, which frequently co-occur with the ground-truth objects. Positive prompts are generated on the basis of ground-truth data. The POPE benchmark covers three datasets, MSCOCO [29], A-OKVQA [40], and GQA [16]. For each dataset, POPE samples 500 images from the validation sets and formulates 6 probing questions (3 positive and 3 negative prompts) for each image and sampling strategy, yielding a total of $27,000$ question-answer pairs. The results are evaluated in terms of Accuracy and F1 Score.

MME: The Multimodal LLM Evaluation (MME) benchmark [2] is another discriminative VQA benchmark, which measures perception and cognition abilities of LVLMs on 14 subtasks comprising $1,193$ images. For each image, there is one positive and one negative question. The evaluation metric is a combined score of the accuracy over all questions and the accuracy+, which is based on each image, that is, both questions need to be answered correctly. Following [25], we average the results over five runs. The standard deviations are given in parentheses.

4.3 Baselines and Implementation Details

We evaluate our proposed ECD method on three state-of-the-art LVLMs, LLaVA 1.5 [31], InstructBLIP [5], and MiniGPT-4 [8] with Vicuna-7B LLM decoder, using nucleus sampling [14] with top_p $= 0.9$. The detailed configuration settings applied in our experiments are summarized in Tab. 1. We compare our approach against regular decoding (denoted as "regular" in our tables) and the contrastive decoding methods VCD [25] and ICD [43]. Throughout our experiments, we use $\alpha = 1$ unless explicitly stated otherwise. All experiments are performed on a single A100 GPU.

Table 1: **LVLM Generation Configurations.** The generation configurations applied in our experiments for nucleus sampling [14] and greedy search.

parameter	nucleus sampling	greedy search
do_sample	True	False
top_p	0.9	1
temperature	1	1
num_beams	1	1
max_length	256	256
min_length	1	1
repetition_penalty	1	1
length_penalty	1	1

Table 2: **Detection Results.** Detection results for the LVLMs LLaVA 1.5 (LV), InstructBLIP (IB), and MiniGPT-4 (MG) with respective hallucination rates CHAIR$_i$ (C_i). The best results in each block are in bold face. The standard deviations are given in parentheses.

LVLM (C_i in %)	Set	ACC ↑ LR	ACC ↑ GB	AUROC ↑ LR	AUROC ↑ GB	AUPRC ↑ LR	AUPRC ↑ GB
LV	[24]	87.1$^{(\pm0.2)}$	87.7$^{(\pm0.3)}$	89.9$^{(\pm0.4)}$	90.8$^{(\pm0.4)}$	68.9$^{(\pm1.1)}$	71.5$^{(\pm0.9)}$
(18.61)	Ours	87.9$^{(\pm0.2)}$	**88.3**$^{(\pm0.3)}$	91.3$^{(\pm0.3)}$	**91.9**$^{(\pm0.4)}$	72.1$^{(\pm1.2)}$	**74.1**$^{(\pm1.0)}$
IB	[24]	91.8$^{(\pm0.1)}$	91.9$^{(\pm0.2)}$	90.2$^{(\pm1.2)}$	90.4$^{(\pm1.1)}$	56.7$^{(\pm5.1)}$	56.7$^{(\pm4.3)}$
(10.13)	Ours	**92.5**$^{(\pm0.1)}$	92.3$^{(\pm0.1)}$	**91.8**$^{(\pm1.0)}$	91.5$^{(\pm0.8)}$	**61.4**$^{(\pm4.4)}$	60.1$^{(\pm4.4)}$
MG	[24]	89.6$^{(\pm0.3)}$	89.7$^{(\pm0.3)}$	88.6$^{(\pm1.8)}$	88.8$^{(\pm1.4)}$	56.6$^{(\pm6.5)}$	56.6$^{(\pm5.9)}$
(13.05)	Ours	90.4$^{(\pm0.2)}$	**90.5**$^{(\pm0.3)}$	90.8$^{(\pm1.6)}$	**90.9**$^{(\pm1.2)}$	**62.5**$^{(\pm8.8)}$	62.2$^{(\pm5.5)}$

5 Results

5.1 Hallucination Detection Results

In this section, we evaluate the information content of our proposed input features (see Sec. 3.2) for probabilistic hallucination detection. The focus of our evaluation is on the AUPRC values as we observe highly imbalanced datasets, i.e., low hallucination rates. The results for the LR and GB classifier are stated in Tab. 2. Our new input features outperform the baseline features in all settings by up to 5.89*pp* in AUPRC values. While the LR and GB classifiers show equal performance on InstructBLIP and MiniGPT-4, we observe superiority of the GB model for LLaVA 1.5 with an improvement of 1.96*pp* in terms of AUPRC. Thus, we employ the GB model in the following experiments.

Table 3: **Discriminative Results on POPE.** Experimental results on the POPE benchmark in terms of accuracy (Acc.), F1 Score and the average inference time per question (time). The best results in each block are in bold face.

			LLaVA 1.5			InstructBLIP			MiniGPT-4		
			time ↓	Acc. ↑	F1 ↑	time ↓	Acc. ↑	F1 ↑	time ↓	Acc. ↑	F1 ↑
MSCOCO	rand.	regular	**0.7**	87.57	87.90	**0.5**	83.90	84.01	**1.3**	54.77	52.67
		VCD	1.3	88.10	88.49	1.0	86.00	85.92	2.6	55.63	51.33
		ICD	1.2	87.70	87.79	0.8	86.67	85.43	1.6	57.57	59.60
		Ours	**0.7**	**89.00**	**89.28**	0.7	**89.00**	**88.69**	1.4	**69.07**	**72.43**
	pop.	regular	**0.6**	83.50	84.57	**0.5**	77.63	79.18	**1.3**	48.60	49.31
		VCD	1.2	85.03	85.87	1.0	78.10	79.46	2.6	50.30	48.50
		ICD	1.2	85.93	86.46	0.8	79.47	78.93	1.6	52.70	57.68
		Ours	0.7	**86.97**	**87.60**	0.7	**82.57**	**83.23**	1.4	**58.63**	**66.27**
	adv.	regular	**0.6**	77.93	80.38	**0.5**	73.90	76.52	**1.3**	47.87	48.96
		VCD	1.2	78.23	80.62	1.0	75.10	77.36	2.5	48.83	47.77
		ICD	1.3	**80.07**	81.53	0.8	77.77	77.77	1.7	52.63	57.90
		Ours	0.7	79.37	**81.58**	0.7	**78.33**	**79.93**	1.4	**57.27**	**65.54**
A-OKVQA	rand.	regular	**0.6**	84.77	86.22	**0.5**	82.03	82.98	**1.3**	49.40	47.40
		VCD	1.2	84.30	85.92	1.0	82.70	83.42	2.6	52.13	48.49
		ICD	1.2	85.20	86.50	0.8	85.57	84.80	1.6	54.83	58.14
		Ours	0.8	**86.50**	**87.83**	0.7	**87.87**	**88.16**	1.4	**65.70**	**70.42**
	pop.	regular	**0.6**	78.17	81.47	**0.5**	75.83	78.35	**1.3**	46.37	47.50
		VCD	1.2	78.50	81.73	1.0	76.83	78.93	2.6	47.00	44.87
		ICD	1.2	**80.20**	82.69	0.8	79.33	79.39	1.6	49.10	55.13
		Ours	0.8	80.03	**82.96**	0.7	**80.23**	**82.13**	1.4	**58.13**	**66.11**
	adv.	regular	**0.6**	68.80	75.25	**0.5**	70.60	74.87	**1.3**	43.90	46.11
		VCD	1.1	69.20	75.65	1.0	70.47	74.69	2.6	45.77	45.31
		ICD	1.2	**71.63**	**76.98**	0.8	72.03	73.90	1.6	45.77	52.74
		Ours	0.8	69.07	75.80	0.7	**72.40**	**76.72**	1.4	**53.70**	**63.57**
GQA	rand.	regular	**0.6**	84.07	85.74	**0.6**	79.97	80.95	**1.3**	50.93	50.10
		VCD	1.2	84.80	86.36	1.0	81.53	82.33	2.5	53.80	53.58
		ICD	1.2	**86.53**	87.65	0.8	83.03	81.84	1.6	55.10	59.05
		Ours	0.8	86.43	**87.78**	0.7	**86.07**	**86.33**	1.4	**64.10**	**69.70**
	pop.	regular	**0.6**	74.60	78.99	**0.6**	73.57	76.34	**1.3**	45.43	47.45
		VCD	1.2	73.40	78.16	1.0	74.13	76.88	2.5	49.33	51.28
		ICD	1.1	**75.73**	**79.72**	0.8	75.93	76.22	1.6	47.83	55.20
		Ours	0.8	74.63	79.14	0.7	**77.63**	**79.77**	1.4	**55.50**	**64.99**
	adv.	regular	**0.6**	69.10	75.50	**0.6**	68.73	73.18	**1.3**	43.97	47.42
		VCD	1.2	69.73	**75.94**	1.0	70.57	74.43	2.6	46.97	49.06
		ICD	1.1	**70.03**	75.89	0.8	71.47	72.84	1.6	47.23	54.91
		Ours	0.8	69.10	75.89	0.7	**72.03**	**76.12**	1.4	**52.80**	**63.80**

Table 4: **Discriminative Results on MME.** Experimental results on the MME benchmark in terms of Perception and Cognition scores [2]. The best results in each block are in bold face.

	LLaVA 1.5		InstructBLIP		MiniGPT-4	
	Perception ↑	Cognition ↑	Perception ↑	Cognition ↑	Perception ↑	Cognition ↑
regular	$1291.1^{(\pm 33.1)}$	$317.4^{(\pm 21.0)}$	$1117.5^{(\pm 23.6)}$	$\mathbf{322.1}^{(\pm 33.3)}$	$409.9^{(\pm 20.2)}$	$163.6^{(\pm 13.0)}$
VCD	$1288.6^{(\pm 33.4)}$	$338.9^{(\pm 16.9)}$	$1155.2^{(\pm 29.9)}$	$291.1^{(\pm 16.4)}$	$355.9^{(\pm 22.2)}$	$140.1^{(\pm 19.5)}$
ICD	$1314.4^{(\pm 27.3)}$	$318.9^{(\pm 22.4)}$	$\mathbf{1258.6}^{(\pm 19.7)}$	$295.4^{(\pm 35.4)}$	$\mathbf{514.2}^{(\pm 29.4)}$	$137.8^{(\pm 13.1)}$
Ours	$\mathbf{1400.3}^{(\pm 13.9)}$	$\mathbf{346.3}^{(\pm 19.2)}$	$1105.2^{(\pm 17.4)}$	$282.3^{(\pm 13.5)}$	$502.8^{(\pm 17.1)}$	$\mathbf{176.4}^{(\pm 25.1)}$

5.2 Discriminative Results

POPE: Tab. 3 summarizes our results on the POPE dataset in terms of accuracy (Acc) and F1 Scores. Our proposed ECD method is superior to the baselines in almost all settings while maintaining low computation costs, improving the F1 Score by up to 23.02*pp*, i.e., 33%. Note that although the probabilistic hallucination detection was trained on the MSCOCO dataset, our results demonstrate a consistent performance improvement across all datasets (MSCOCO, A-OKVQA, and GQA) underlining the ability of the meta classifier to judge hallucinations on new data. Furthermore, we observe consistent performance across all sampling strategies (random, popular, and adversarial) showing that the meta classifier effectively learned hallucinatory concepts beyond the language bias induced by the LVLM training [35]. For a detailed analysis of the precision and recall values, we refer to supplementary material B, unveiling the outstanding ability of our method to accurately negate negative prompts, which contain hallucinations.

MME: The MME benchmark evaluates hallucinations beyond the object level and measures general perception and cognition abilities. Tab. 4 presents our results. For LLaVA 1.5 and MiniGPT-4, our method not only improves the perception ability, but also enhances the performance in cognition and reasoning tasks compared to the baseline methods. Moreover, the detailed evaluation of the 14 subtasks and computational time in supplementary material B shows that ECD outperforms the baselines on most of these individual tasks and demonstrates the superior performance-cost trade-off, that is, ECD outperforms the baselines with respect to performance and inference time. Note that although the averaged ECD perception and cognition scores for InstructBLIP are below the baseline scores, our analysis of the subtasks shows that ECD outperforms the baselines on individual tasks while maintaining low computational costs.

Table 5: **Generative Results.** Experimental results on the CHAIR benchmark for the open-ended captioning tasks using the MSCOCO and AMBER datasets. The results are stated in terms of average inference time per image caption (time), CHAIR_i (C_i), CHAIR_s (C_s), and Coverage (Cov.). The best results in each block are in bold face.

	MSCOCO [29]											
	LLaVA 1.5				InstructBLIP				MiniGPT-4			
	time ↓	C_i ↓	C_s ↓	Cov. ↑	time ↓	C_i ↓	C_s ↓	Cov. ↑	time ↓	C_i ↓	C_s ↓	Cov. ↑
regular	**2.7**	17.86	55.00	82.14	**3.5**	9.26	30.80	90.74	**10.1**	11.60	**27.47**	88.40
VCD	5.3	16.32	53.80	83.68	6.3	8.45	30.80	91.55	18.9	10.58	28.60	89.42
ICD	4.5	14.27	45.40	85.73	5.8	10.92	37.80	89.08	15.4	10.51	28.20	89.49
Ours	3.6	**12.12**	**43.40**	**87.88**	4.6	**7.28**	**26.60**	**92.72**	12.8	**9.25**	31.66	**90.75**
	AMBER [21]											
	LLaVA 1.5				InstructBLIP				MiniGPT-4			
	time ↓	C_i ↓	C_s ↓	Cov. ↑	time ↓	C_i ↓	C_s ↓	Cov. ↑	time ↓	C_i ↓	C_s ↓	Cov. ↑
regular	**2.2**	9.99	45.56	**51.97**	**2.9**	8.74	39.38	**51.56**	**9.4**	16.68	61.78	57.71
VCD	4.1	8.16	38.42	51.63	5.4	7.73	34.36	50.50	17.4	14.18	**53.09**	58.23
ICD	3.5	8.46	36.87	49.98	5.3	8.11	35.71	49.02	15.3	16.83	58.11	55.35
Ours	2.8	**7.04**	**33.20**	51.21	4.2	**6.00**	**26.45**	50.39	12.4	**13.14**	61.58	**61.03**

5.3 Generative Results

MSCOCO: In addition to the discriminative results, we also evaluate our method on the open-ended captioning task. Note that for MiniGPT-4, we apply the parameter $\alpha = 6$ (see Sec. 5.4 for an ablation study for α). The results are summarized in Tab. 5. ECD distinctly reduces the hallucination rate both at the instance and sentence level, while simultaneously increasing the detailedness of the generated response in terms of Coverage. In all experiments, ECD is superior to the baseline methods VCD and ICD with respect to performance and computational time. Only in the case of MiniGPT-4, ECD increases the sentence level hallucination rate CHAIR_s while still decreasing the total number of hallucinations CHAIR_i and simultaneously increasing the Coverage. Compared to regular decoding, ECD reduces the instance level hallucination rate by up to $5.74pp$, i.e., 32% while at the same time increasing the Coverage by $5.74pp$, i.e., 7% while maintaining low computational costs.

AMBER: Since the ECD meta classifier was trained on the MSCOCO dataset, we investigate the potential of our classifier on new concepts. The results in Tab. 5 underline our findings from the discriminative results (Sec. 5.2). Again, ECD successfully suppresses hallucinations due to the classifier's ability to judge hallucinations on new data. More precisely, we reduce the instance level hallucination rate by $2.95pp$, i.e., 30% while maintaining the detailedness of the gen-

Table 6: **Ablation: Decoding Configuration.** Ablation results on the discriminative POPE benchmark [28] for nucleus sampling [14] decoding with top_p $= 1$ and greedy decoding. The best results in each block are in bold face.

		LLaVA 1.5			InstructBLIP			MiniGPT-4		
		time ↓	Acc. ↑	F1 ↑	time ↓	Acc. ↑	F1 ↑	time ↓	Acc. ↑	F1 ↑
$p=1$	regular	**0.6**	82.47	83.55	**0.5**	74.53	75.97	**1.3**	45.37	44.87
	VCD	1.2	84.13	85.13	1.0	77.53	79.13	2.6	48.33	46.40
	ICD	1.2	85.70	86.22	0.8	79.80	79.37	1.7	52.57	57.07
	Ours	0.8	**86.20**	**86.85**	0.6	**82.07**	**82.87**	1.4	**58.07**	**66.26**
greedy	regular	**0.6**	**87.70**	88.21	**0.5**	82.40	83.36	**1.3**	**71.20**	**73.35**
	VCD	1.2	**87.70**	88.21	0.9	81.03	81.99	2.6	68.60	68.11
	ICD	1.2	88.10	**88.31**	0.8	81.87	81.10	1.7	65.80	70.55
	Ours	0.8	87.60	88.13	0.7	**83.27**	**83.83**	1.4	60.17	69.98

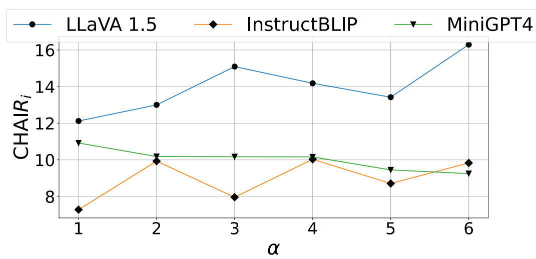


Fig. 2: Ablation study for hyperparameter α.

erated response and low computational costs. While we see a minor decrease in Coverage for LLaVA 1.5 and InstructBLIP, ECD effectively reduces the hallucination rate in terms of both, CHAIR$_i$ and CHAIR$_s$. In the case of MiniGPT-4, we again observe an increase in CHAIR$_s$ while effectively improving the instance level hallucination rate CHAIR$_i$ as well as the Coverage.

5.4 Ablations

LVLM Decoding Configuration: We conduct additional experiments for the POPE MSCOCO popular setting using nucleus sampling with top_p $= 1$ and greedy search decoding. The results are summarized in Tab. 6. While the results for regular nucleus sampling with top_p $= 1$ are below the top_p $= 0.9$ results from Tab. 3, all contrastive decoding strategies maintain the performance in the top_p $= 1$ setting. Again, our method outperforms VCD [25] and ICD [43] with respect to performance and computational time. Note that for the greedy search setting, the contrastive decoding methods achieve a minor performance increase only, where for MiniGPT-4 regular decoding performs best.

ECD Hyperparameter: Moreover, we investigate the influence of the hyperparameter α in our method on the CHAIR open-ended text generation task. Fig. 2 depicts the results. While for LLaVA 1.5 and InstructBLIP the best results are achieved applying $\alpha = 1$, for MiniGPT-4 the best performance is achieved with $\alpha = 6$, i.e., a higher magnitude of hallucination correction. However, note that for MiniGPT-4 different α values result in minor performance changes in terms of the hallucination rate CHAIR_i only.

6 Limitations

The focus of our paper is on visual hallucinations of LVLMs, where contextual hallucinations in LLMs might have different origins, which need to be studied to ensure a successful transfer of our method to the unimodal domain. However, note that many hallucination features are specifically designed for the transformer architecture, which can be directly transferred to LLMs, further broadening the impact of this work. Moreover, recent advances in video LVLMs motivate the investigation of temporal hallucinations, a problem we will tackle in future work.

7 Conclusion

In this paper, we investigate the power of probabilistic hallucination detection for contrastive decoding. We introduce Efficient Contrastive Decoding (ECD), a lightweight and training-free method, which shifts the LVLM output distribution towards accurate responses during decoding by penalizing hallucinations. Extensive experimental results demonstrate the efficacy of our proposed method, which outperforms state-of-the-art methods on various LVLM baselines. Our experiments show that ECD not only mitigates hallucinations but also enhances the perception capabilities of LVLMs. Moreover, in contrast to existing methods, our lightweight approach is computationally efficient, adding only minor computational overhead to the decoding process.

Ethics Considerations Our work addresses the hallucination issue in state-of-the-art LVLMs enhancing the reliability and integrity of LVLMs in real-world scenarios, especially in safety-critical applications such as autonomous driving or medicine. Moreover, our work does not include any personal data, human subjects or sensitive data.

Disclaimer The results, opinions and conclusions expressed in this publication are not necessarily those of Volkswagen Aktiengesellschaft.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Azaria, A., Mitchell, T.: The internal state of an llm knows when it is lying. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 967–976 (2023)
2. Chaoyou Fu and Peixian Chen and Yunhang Shen and Yulei Qin and Mengdan Zhang and Xu Lin and Jinrui Yang and Xiawu Zheng and Ke Li and Xing Sun and Yunsheng Wu and Rongrong Ji: Mme: A comprehensive evaluation benchmark for multimodal large language models (2024), `https://arxiv.org/abs/2306.13394`
3. Chen, T., Navratil, J., Iyengar, V., Shanmugam, K.: Confidence scoring using whitebox meta-models with linear classifier probes. In: Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics. pp. 1467–1475 (2019)
4. Chen, Z., Zhao, Z., Luo, H., Yao, H., Li, B., Zhou, J.: Halc: Object hallucination reduction via adaptive focal-contrast decoding. In: Proceedings of the 41st International Conference on Machine Learning (2024)
5. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (2023)
6. Dan Hendrycks and Kevin Gimpel: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017)
7. Davis, J., Goadrich, M.: The relationship between precision-recall and roc curves. In: Proceedings of the 23rd International Conference on Machine Learning. pp. 233–240 (2006)
8. Deyao Zhu and Jun Chen and Xiaoqian Shen and Xiang Li and Mohamed Elhoseiny: MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In: The Twelfth International Conference on Learning Representations (2024)
9. Fieback, L., Dash, B., Spiegelberg, J., Gottschalk, H.: Temporal performance prediction for deep convolutional long short-term memory networks. In: Advanced Analytics and Learning on Temporal Data. pp. 145–158 (2023)
10. Freitag, Markus and Al-Onaizan, Yaser: Beam search strategies for neural machine translation. In: Proceedings of the First Workshop on Neural Machine Translation (2017)
11. Fuxiao Liu and Kevin Lin and Linjie Li and Jianfeng Wang and Yaser Yacoob and Lijuan Wang: Mitigating hallucination in large multi-modal models via robust instruction tuning. In: The Twelfth International Conference on Learning Representations (2024)
12. Gunjal, A., Yin, J., Bas, E.: Detecting and preventing hallucinations in large vision language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 18135–18143 (2024)
13. Haoxiang Gao and Zhongruo Wang and Yaqian Li and Kaiwen Long and Ming Yang and Yiqing Shen: A survey for foundation models in autonomous driving (2024), `https://arxiv.org/abs/2402.01105`
14. Holtzman, A., Buys, J., Du, L., Forbes, M., Choi, Y.: The curious case of neural text degeneration. In: International Conference on Learning Representations (2020)
15. Hong Yan, Lijun Liu, Xupeng Feng & Qingsong Huang: Overcoming language priors with self-contrastive learning for visual question answering. In: Multimedia Tools and Applications. pp. 16343–16358 (2023)

16. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6693–6702 (2019)
17. Jiang, C., Xu, H., Dong, M., Chen, J., Ye, W., Yan, M., Ye, Q., Zhang, J., Huang, F., Zhang, S.: Hallucination augmented contrastive learning for multimodal large language model. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27026–27036 (2024)
18. Jiang, Y., A. Omiye, J., Zakka, C., Moor, M., Gui, H., Alipour, S., Shahabeddin Mousavi, S., Chen, J.H., Rajpurkar, P., Daneshjou, R.: Evaluating general vision-language models for clinical medicine. medRxiv (2024)
19. Jing, L., Li, R., Chen, Y., Du, X.: Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 5042–5063 (2024)
20. Junyang Wang and Yiyang Zhou and Guohai Xu and Pengcheng Shi and Chenlin Zhao and Haiyang Xu and Qinghao Ye and Ming Yan and Ji Zhang and Jihua Zhu and Jitao Sang and Haoyu Tang: Evaluation and analysis of hallucination in large vision-language models (2023), `https://arxiv.org/abs/2308.15126`
21. Junyang Wang and Yuhang Wang and Guohai Xu and Jing Zhang and Yukai Gu and Haitao Jia and Jiaqi Wang and Haiyang Xu and Ming Yan and Ji Zhang and Jitao Sang: Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation (2024), `https://arxiv.org/abs/2311.07397`
22. Kamil Kowol. and Matthias Rottmann. and Stefan Bracke. and Hanno Gottschalk: Yodar: Uncertainty-based sensor fusion for vehicle detection with camera and radar sensors. In: Proceedings of the 13th International Conference on Agents and Artificial Intelligence. pp. 177–186 (2021)
23. Kim, M., Kim, M., Bae, J., Choi, S., Kim, S., Chang, B.: Exploiting semantic reconstruction to mitigate hallucinations in vision-language models. In: Computer Vision – ECCV 2024. pp. 236–252 (2025)
24. Laura Fieback, Jakob Spiegelberg, Hanno Gottschalk: Metatoken: Detecting hallucination in image descriptions by meta classification. In: Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. pp. 126–137 (2025)
25. Leng, Sicong and Zhang, Hang and Chen, Guanzheng and Li, Xin and Lu, Shijian and Miao, Chunyan and Bing, Lidong: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13872–13882 (2024)
26. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (2023)
27. Li, X.L., Holtzman, A., Fried, D., Liang, P., Eisner, J., Hashimoto, T., Zettlemoyer, L., Lewis, M.: Contrastive decoding: Open-ended text generation as optimization. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. pp. 12286–12312 (2023)
28. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305 (2023)
29. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision – ECCV 2014. pp. 740–755 (2014)

30. Linxi Zhao and Yihe Deng and Weitong Zhang and Quanquan Gu: Mitigating object hallucination in large vision-language models via classifier-free guidance (2024), `https://arxiv.org/abs/2402.08680`
31. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems (2023)
32. Maag, K., Rottmann, M., Gottschalk, H.: In: 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI). pp. 502–509 (2020)
33. Maag, K., Rottmann, M., Varghese, S., Hüger, F., Schlicht, P., Gottschalk, H.: Improving video instance segmentation by light-weight temporal uncertainty estimates. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2021)
34. Maha Elbayad and Jiatao Gu and Edouard Grave and Michael Auli: Depth-adaptive transformer. In: International Conference on Learning Representations (2020)
35. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4035–4045 (2018)
36. Rottmann, M., Colling, P., Paul Hack, T., Chan, R., Hüger, F., Schlicht, P., Gottschalk, H.: Prediction error meta classification in semantic segmentation: Detection via aggregated dispersion measures of softmax probabilities. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–9 (2020)
37. S. Kullback and R. A. Leibler: On information and sufficiency. The Annals of Mathematical Statistics pp. 79–86 (1951)
38. Schubert, M., Kahl, K., Rottmann, M.: Metadetect: Uncertainty quantification and prediction quality estimates for object detection. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–10 (2021)
39. Schuster, T., Fisch, A., Gupta, J., Dehghani, M., Bahri, D., Tran, V.Q., Tay, Y., Metzler, D.: Confident adaptive language modeling. In: Proceedings of the 36th International Conference on Neural Information Processing Systems (2022)
40. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: Computer Vision – ECCV 2022. pp. 146–162 (2022)
41. Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, Enhong Chen: Woodpecker: hallucination correction for multimodal large language models. Sci. China Inf. Sci. 67 (2024)
42. Teerapittayanon, S., McDanel, B., Kung, H.T.: Branchynet: Fast inference via early exiting from deep neural networks. In: 2016 23rd International Conference on Pattern Recognition (ICPR). pp. 2464–2469 (2016)
43. Wang, X., Pan, J., Ding, L., Biemann, C.: Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 15840–15853 (2024)
44. Wu, J., Liu, Q., Wang, D., Zhang, J., Wu, S., Wang, L., Tan, T.: Logical closed loop: Uncovering object hallucinations in large vision-language models. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 6944–6962 (2024)
45. Xiaoyu Tian and Junru Gu and Bailin Li and Yicheng Liu and Yang Wang and Zhiyong Zhao and Kun Zhan and Peng Jia and XianPeng Lang and Hang Zhao: DriveVLM: The convergence of autonomous driving and large vision-language models. In: 8th Annual Conference on Robot Learning (2024)
46. Yiyang Zhou and Chenhang Cui and Jaehong Yoon and Linjun Zhang and Zhun Deng and Chelsea Finn and Mohit Bansal and Huaxiu Yao: Analyzing and miti-

gating object hallucination in large vision-language models. In: The Twelfth International Conference on Learning Representations (2024)
47. Yung-Sung Chuang and Yujia Xie and Hongyin Luo and Yoon Kim and James R. Glass and Pengcheng He: Dola: Decoding by contrasting layers improves factuality in large language models. In: The Twelfth International Conference on Learning Representations (2024)