Activation Steering Decoding: Mitigating Hallucination in Large Vision-Language Models through Bidirectional Hidden State Intervention

Jingran Su[*1], **Jingfan Chen**[*1], **Hongxin Li**[*2,3,5], **Yuntao Chen**[†4]
Qing Li[†1], **Zhaoxiang Zhang**[†2,3,5,6]
[1]The Hong Kong Polytechnic University
[2]New Laboratory of Pattern Recognition, CASIA
[3]State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA
[4]Hong Kong Institute of Science & Innovation, CASIA
[5]University of Chinese Academy of Sciences [6]Shanghai Artificial Intelligence Laboratory

Abstract

Large Vision-Language Models (LVLMs) have demonstrated impressive capabilities in multimodal understanding, but they frequently suffer from hallucination - generating content inconsistent with visual inputs. In this work, we explore a novel perspective on hallucination mitigation by examining the intermediate activations of LVLMs during generation. Our investigation reveals that hallucinated content manifests as distinct, identifiable patterns in the model's hidden state space. Motivated by this finding, we propose **A**ctivation **S**teering **D**ecoding (ASD), a training-free approach that mitigates hallucination through targeted intervention in the model's intermediate activations. ASD operates by first identifying directional patterns of hallucination in the activation space using a small calibration set, then employing a contrast decoding mechanism that computes the difference between positive and negative steering predictions. This approach effectively suppresses hallucination patterns while preserving the model's general capabilities. Extensive experiments demonstrate that our method significantly reduces hallucination across multiple benchmarks while maintaining performance on general visual understanding tasks. Notably, our approach requires no model re-training or architectural modifications, making it readily applicable to existing deployed models.

1 Introduction

Large Vision Language Models (LVLMs), while demonstrating impressive capabilities, struggle with a fundamental issue known as *hallucination* where generated textual descriptions fail to align accurately with visual semantics (Liu et al., 2024a; Zhai et al., 2023; Zhao et al., 2023). These failures not only degrade the performance of LVLMs in practical scenarios but also undermine their credibility in high-stakes applications like medical imaging, autonomous driving, and legal systems (Wang, 2024; Magesh et al., 2024).

While existing approaches mitigate hallucination through enhanced data quality (Liu et al., 2023a; Yu et al., 2024a) and carefully designed training objectives (Chen et al., 2023; Jiang et al., 2024; Yue et al., 2024), such post-training solutions may present challenges for real-world deployments where models need to adapt rapidly to scenarios with minimal computational overhead and maximum flexibility.

Recent attempts have made significant progress in exploring training-free solutions as crucial alternatives. These approaches can be broadly categorized into module-level methods (Zhao et al., 2024; Deng et al., 2024; Yu et al., 2025; An et al., 2024) that leverage richer visual modules, and logit-level methods (Leng et al., 2024; Zhu et al., 2024) that reduce the model's reliance on language priors or statistical biases. Both approaches share a fundamental principle: strengthening visual evidence through either enhanced visual signals or additional visual cues during the inference process.

While these approaches provide valuable insights, they focus on specific assumptions (e.g., attention loss in image regions). In contrast, this work aims to address this in a more fundamental way. We propose an approach by directly steering the model with a hallucination-aware distributional indicator to generate hallucination-free descriptions. We first analyze hallucination behavior in LVLMs by examining intermediate activation, i.e. hidden state[1], distributions. Our empirical investigation reveals that hallucinated content manifests as distinct, identifiable patterns in the model's intermediate activation. Building on this insight and to achieve effective steering, we propose **A**ctivation

[*] Equal contribution.

[†] Equally advising corresponding authors. E-mails: zhaoxiang.zhang@ia.ac.cn, csqli@comp.polyu.edu.cn, chenyuntao08@gmail.com

[1] In this paper, we do not differentiate between the terms "hidden state" and "intermediate activation", treating them as interchangeable concepts.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12964–12974
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

Steering **D**ecoding (ASD), a training-free approach that directly intervenes in the model's intermediate activations to mitigate hallucination.

Our method operates by first identifying the directional patterns of hallucination in the intermediate activation space using a small calibration set, then employing a contrast decoding mechanism that computes the difference between positive and negative steering predictions. Extensive experiments demonstrate that our method achieves substantial reductions in hallucination rates (over 10.0% improvement on CHAIR and over 10% F1 score improvement on POPE) while maintaining or even enhancing performance on general visual understanding tasks. Notably, our method requires no re-training or architectural modifications, making it readily applicable to deployed models.

The main contributions of this paper include: 1) a systematic empirical study that reveals the distinct patterns of hallucination in LVLMs intermediate activation space, providing insights into the internal mechanisms of LVLMs; 2) ASD: a novel, training-free method for hallucination reduction through targeted intervention in intermediate activations; 3) comprehensive empirical evaluation demonstrating significant reduction in hallucination across diverse scenarios while maintaining model performance on standard tasks.

2 Related Works

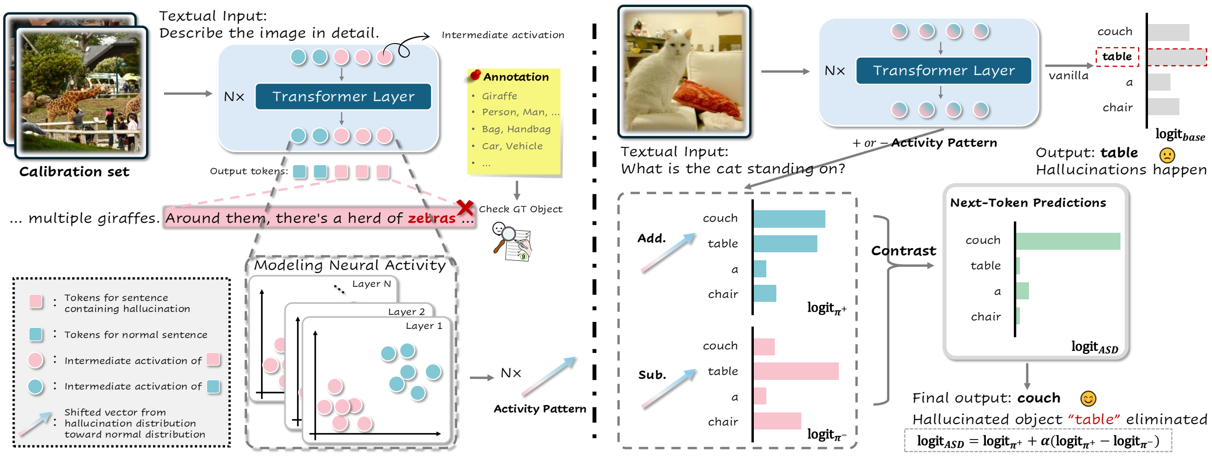
Hallucination in LVLMs. Hallucination was initially studied and defined in the context of language models, describing outputs that deviate from factual or contextual information. In LVLMs, hallucination specifically refers to model outputs that are inconsistent with the input visual information. To address this challenge, various approaches have been proposed. Some works enhance visual features through diverse visual encoders or visual tools (Jain et al., 2024; He et al., 2024; Jiao et al., 2024), and employ specialized modules to control cross-modal alignment (Zhai et al., 2023). Other researchers have approached this problem from a data-centric perspective, introducing contrastive examples and adversarial samples to increase training data diversity (Liu et al., 2023a; Yu et al., 2024a), while also implementing denoising and regeneration strategies to improve overall data quality (Wang et al., 2024; Yue et al., 2024). Additional works have incorporated extra supervision signals during training to strengthen visual feature representations (Chen et al., 2023; Jiang et al., 2024; Yue et al., 2024), and some have employed reinforcement learning techniques to suppress model hallucination (Zhao et al., 2023; Zhou et al., 2024; Sun et al., 2023; Yu et al., 2024b). However, these methods either require substantial additional data or involve expensive training processes. Furthermore, several training-free methods have been proposed. These include interventions in the model's output process through contrast decoding (Leng et al., 2024; Zhu et al., 2024), guidance from auxiliary models (Zhao et al., 2024; Deng et al., 2024; Yu et al., 2025; An et al., 2024), and post-processing techniques to eliminate hallucinated content from the outputs (Yin et al., 2023; Lee et al., 2023; Zhou et al., 2023).

Activation Steering. Our method analyzes and intervenes in the model's representation space, which relates to the recent technique of activation steering (or representation engineering) in language models (Subramani et al., 2022; Turner et al., 2023; Jorgensen et al., 2023; Panickssery et al., 2023; Liu et al., 2023b; Zou et al., 2023). Activation steering is a technique used to guide model behavior by manipulating neuron activations. Most relevant to our work are several studies (Panickssery et al., 2023; Turner et al., 2023), where they use semantically opposite prompt pairs (such as the prompts "Love" and "Hate") to generate steering vectors that, when added to model activations, can control model behavior. Different from these approaches, our approach identifies hallucination-specific patterns through analysis of activations rather than prompt engineering, and presents a contrast decoding mechanism that enables robust hallucination mitigation while maintaining generation quality.

3 Preliminary

This section introduces the key notations used throughout this paper. Consider a LVLM $\pi(\cdot)$ that accepts image v and language x inputs to generate text sequences $\mathbf{y} = (y_1, ..., y_n)$. As the inputs pass through the model's transformer architecture, it generates a series of intermediate activations $\mathbf{Z} = \mathbf{z}_1, ..., \mathbf{z}_L$ at each layer l, with $\mathbf{z}_l \in \mathbb{R}^d$. The model generates each token through sampling from the following distribution:

$$\begin{aligned} y_t &\sim \pi(y_t|x, v, y_{<t}), \\ &\propto \exp(\mathrm{logit}_\pi(y_t|x, v, y_{<t})), \end{aligned}$$



(a) Hallucination Feature Collecting and Modeling

(b) Activation Steering Decoding at Inference Time

Figure 1: Overview of our proposed method. **Left:** The token-level hallucination feature collection process, where we extract hidden states from the model and annotate them based on whether they belong to sentences containing hallucinated objects (not present in the ground truth). The steering vector is computed as the difference between mean hidden states of hallucinated and non-hallucinated tokens. **Right:** Illustration of Activation Steering Decoding, which performs two forward passes with opposite steering directions and contrasts their logits to obtain the final output distribution, effectively suppressing hallucination patterns while preserving semantic information.

where $\text{logit}_\pi(y_t|\cdot)$ represents the unnormalized log probabilities for token y_t.

4 How Do Hidden States Differ during Hallucination?

We start by analyzing how hallucinations manifest in the hidden states of LVLMs during generation. We hypothesize that hallucinated content exhibits distinct patterns in the model's hidden state space compared to factual generations. To investigate this hypothesis, we propose a framework designed to systematically extract the model's hidden representations paired with labels indicating hallucination occurrences in Sec. 4.1 and analyze their corresponding hidden state representations via linear probing in Sec. 4.2.

4.1 A Framework for Representation Collection

To systematically investigate hallucination patterns in the given base model π_{base}, we develop a scalable framework for collecting paired hidden states and hallucination labels for it. Our approach focuses specifically on object hallucination, a well-defined and measurable form of multimodal hallucination that occurs when a model generates references to objects not present in the input image. The following details our data collection process:

Image-Description Pair Generation. We utilize the MSCOCO dataset (Lin et al., 2014) as our primary data source due to its rich annotations for segmentation and diverse visual content. For each image v_i in the dataset, we query the base mode π_{base} with prompt x = "Please describe the image in detail." to generate a detailed description $\mathbf{y}_i$.

The generated description $\mathbf{y}_i$ reflects the model's intrinsic perception of the input image v_i, which may contain hallucinated content that deviates from the actual visual information.

Activation Collection and Annotation. $\mathcal{O} = \{o_1, o_2, ..., o_{80}\}$ represent the set of 80 predefined object categories in the MSCOCO dataset. For each object category o, we collect a set of synonyms $\mathcal{C}(o)$ to ensure comprehensive object extraction. Each image v_i is associated with its ground truth object set $G(v_i) \subseteq \mathcal{O}$ based on MSCOCO annotations. For each generated description $\mathbf{y}_i$, we employ the Natural Language Toolkit library to segment it into individual sentences $\{\mathbf{s}_{i,1}, \mathbf{s}_{i,2}, \ldots, \mathbf{s}_{i,j}\}$, where each $\mathbf{s}_{i,j}$ is a subsequence of tokens representing a single sentence:

$$\mathbf{s}_{i,j} = (y_1^{i,j}, y_2^{i,j}, \ldots, y_p^{i,j}), \quad \text{with} \bigcup_j \mathbf{s}_{i,j} = \mathbf{y}_i.$$

We then identify all mentioned objects $O(\mathbf{s}_{i,j})$ in the sentence $\mathbf{s}_{i,j}$ by:

$$O(\mathbf{s}_{i,j}) = \{o \in \mathcal{O} \mid \begin{matrix} \text{substr}(o, \mathbf{s}_{i,j}), \text{ or} \\ \exists c \in \mathcal{C}(o), \text{substr}(c, \mathbf{s}_{i,j}) \end{matrix}\},$$

$$\text{substr}(x, y) \iff x \text{ is a substring of } y.$$

We define the hallucination label $L(y_p^{i,j})$ for a token $y_p^{i,j} \in \mathbf{s}_{i,j}$ based on whether the sentence $\mathbf{s}_{i,j}$

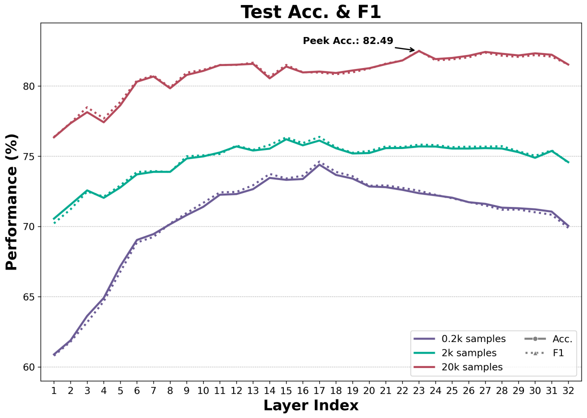


Figure 2: Test accuracy and F1 scores for hallucination versus non-hallucination classification across different layers of LLaVA-1.5-7B with varying training sample sizes (0.2k, 2k, and 20k).

includes any non-existent objects. Mathematically:

$$L(y_p^{i,j}) = \begin{cases} 1 & \text{if } O(\mathbf{s}_{i,j}) \setminus G(I_i) \neq \varnothing, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathbf{Z}(y)$ indicates the hidden state of all layer for token y. The final dataset of paired activations and hallucination labels is constructed as: $\bigcup_i \left(\mathbf{Z}(y_p^{i,j}), L(y_p^{i,j})\right)$.

4.2 Linear Probing of Hidden States

To investigate the patterns of hidden states when occurring hallucination, we perform linear probing of LLaVA1.5-7B across its entire architecture. Specifically, we randomly sample 500 images from the MSCOCO training set and employ the methodology described in Sec. 4.1 to extract hidden state representations across all 32 transformer layers. This initial collection yields an imbalanced dataset comprising 42,160 non-hallucinated samples and 12,113 hallucinated samples. We then construct a balanced dataset by randomly sampling 11,000 instances from each class, resulting in a final dataset of 22,000 samples. We reserve 2,000 samples as a held-out test set and use the remaining 20,000 for training. We conduct a series of linear probing experiments with varying amounts of training data, independently training linear classifiers for each of the 32 layers' hidden states to track how hallucination-related information is encoded across the model's layers.

Fig. 2 presents the accuracy and F1 scores across model layers under varying training set sizes. Our analysis reveals several significant findings. First, the amount of training data exhibits a substantial impact on the classifier's discriminative capability, with approximately 20k samples being necessary to establish reliable patterns. This suggests that hallucination signatures, while consistent, require sufficient data to be accurately characterized. Moreover, we observe that hidden states in the middle and latter layers demonstrate superior representational power for hallucination detection, indicating a progressive accumulation of hallucination-relevant features across the model's hierarchy. Most notably, the probing performance reveals that hallucination-related information is remarkably well-preserved and linearly separable in the hidden state space, achieving probing accuracy of 82.49% in the middle layers with just 20k training tokens. This pronounced linear separability provides compelling evidence that hallucinated content manifests as distinct, consistent patterns in the model's hidden state space, which in turn suggests that targeted intervention at the hidden state level could effectively mitigate hallucination behavior.

5 Activation Steering Decoding

Motivated by our empirical findings that hallucination patterns are distinctly encoded and linearly separable in the model's hidden states, we propose Activation Steering Decoding, a novel decoding strategy that directly intervenes in the model's hidden activations to mitigate hallucination.

Steering Vector Modeling. Given the paired data $\bigcup_i \{\left(\mathbf{Z}(y_p^{i,j}), L(y_p^{i,j})\right)\}$ collected from Sec. 4.1, we calculate a steering vector that captures the direction from hallucination to non-hallucination in the hidden state space. For each layer l, we compute the difference between mean activations of non-hallucinated and hallucinated tokens:

$$\mathbf{v}^l = \frac{1}{P} \sum_{L(y)=1} \mathbf{z}_l(y) - \frac{1}{N} \sum_{L(y)=0} \mathbf{z}_l(y), \qquad (1)$$

where P and N are the numbers of factual and hallucinated tokens respectively.

Steering Vector Injection. The most straightforward approach to leveraging the extracted steering vectors is directly intervening in the hidden states:

$$\mathbf{z}_l^{\text{steered}} = \mathbf{z}_l + \lambda \mathbf{v}_l, \qquad (2)$$

where λ regulates the steering strength. While this approach effectively reduces hallucination as λ increases , it risks distorting the semantic information encoded in the hidden states (see ablation studies in Sec. 6.5.3).

Method	MSCOCO		A-OKVQA		GQA	
	%Accuracy	%F1 Score	%Accuracy	%F1 Score	%Accuracy	%F1 Score
Greedy Decoding						
LLaVA1.5-7B	85.13 ↑0.00	86.03 ↑0.00	78.99 ↑0.00	82.61 ↑0.00	76.60 ↑0.00	80.98 ↑0.00
+ VCD	85.16 ↑0.03	86.04 ↑0.01	78.92 ↓0.07	82.58 ↓0.03	76.49 ↓0.11	80.94 ↓0.04
+ VDD-None	86.87 ↑1.74	87.26 ↑1.23	82.02 ↑3.01	84.57 ↑1.96	79.99 ↑3.39	83.04 ↑2.06
+ ASD (Ours)	**88.01 ↑2.88**	**87.87 ↑1.84**	**85.10 ↑6.11**	**85.65 ↑3.04**	**83.49 ↑6.89**	**83.98 ↑3.00**
Qwen-VL-Chat	86.44 ↑0.00	86.12 ↑0.00	85.92 ↑0.00	85.80 ↑0.00	75.23 ↑0.00	67.70 ↑0.00
+ VCD	86.42 ↓0.02	86.31 ↑0.19	85.64 ↓0.28	85.70 ↓0.10	77.06 ↑1.83	71.19 ↑3.49
+ VDD-None	86.72 ↑0.28	86.45 ↑0.33	85.58 ↓0.34	85.58 ↓0.22	75.88 ↑0.65	68.94 ↑1.24
+ ASD (Ours)	**88.09 ↑1.65**	**87.96 ↑1.84**	**87.29 ↑1.37**	**87.29 ↑1.49**	**83.77 ↑8.54**	**82.21 ↑14.51**
Direct Sampling						
LLaVA1.5-7B	81.49 ↑0.00	82.93 ↑0.00	75.97 ↑0.00	80.04 ↑0.00	73.71 ↑0.00	78.48 ↑0.00
+ VCD	85.41 ↑3.92	86.27 ↑3.34	78.87 ↑2.90	82.55 ↑2.51	76.53 ↑2.82	80.97 ↑2.49
+ VDD-None	85.77 ↑4.28	86.28 ↑3.35	81.02 ↑5.05	83.73 ↑3.69	79.41 ↑5.70	82.45 ↑3.97
+ ASD (Ours)	**87.19 ↑5.70**	**87.15 ↑4.22**	**84.63 ↑8.66**	**85.34 ↑5.30**	**83.19 ↑9.48**	**83.89 ↑5.41**
Qwen-VL-Chat	84.16 ↑0.00	83.59 ↑0.00	83.01 ↑0.00	82.79 ↑0.00	74.54 ↑0.00	67.12 ↑0.00
+ VCD	86.47 ↑2.31	86.24 ↑2.65	85.52 ↑2.51	**85.60 ↑2.81**	77.42 ↑2.88	71.83 ↑4.71
+ VDD-None	86.10 ↑1.94	85.78 ↑2.19	84.96 ↑1.95	84.99 ↑2.20	75.71 ↑1.17	68.68 ↑1.56
+ ASD (Ours)	**87.03 ↑2.87**	**86.86 ↑3.27**	**85.69 ↑2.68**	85.52 ↑2.73	**82.84 ↑8.30**	**80.77 ↑13.65**

Table 1: Performance evaluation of our method against baselines and related approaches on POPE benchmark under two decoding strategies: Greedy Decoding and Direct Sampling. The base models (LLaVA1.5-7B and Qwen-VL-Chat) are compared with VCD and VDD-None (existing methods) as well as our proposed approach. Results are reported in terms of Accuracy (%) and F1 Score (%). The proposed method achieves consistent and notable improvements over all baselines and related methods, with the best results highlighted in bold.

Activation Steering Decoding. To achieve more stable hallucination reduction while preserving generation quality, we propose Activation Steering Decoding. Let π^+ and π^- denote the model under positive (i.e., $\lambda > 0$) and negative (i.e., $\lambda < 0$) steering using Eq. (2) respectively, applying the same steering vector in opposite directions. The final logits for next token prediction are obtained through following:

$$\text{logit}_{ASD} = (1+\alpha)\cdot \text{logit}_{\pi^+} - \alpha \cdot \text{logit}_{\pi^-}, \quad (3)$$

where α is the contrastive weight coefficient. This contrast mechanism is effective because the difference operation amplifies our steering's impact on output logits, while allowing us to use a relatively small steering intensity to better preserve semantic integrity in the hidden states. This property makes our approach more robust and less likely to disturb the model's normal generation process compared to direct steering.

6 Experiments

In this section, we evaluate our proposed Activation Steering Decoding method on various multimodal benchmarks. Our experiments aim to assess both hallucination reduction and general visual comprehension capabilities.

6.1 Benchmarks

We conduct experiments on two categories of benchmarks:

Visual Hallucination. POPE evaluates object hallucination through yes/no questions about object presence. It contains 27,000 question-answer pairs sourced equally from MS-COCO, A-OKVQA, and GQA datasets (9,000 each). The questions are categorized into three types Random, Popular, and Adversarial. **CHAIR** measures object hallucination in image captioning tasks. It provides fine-grained annotations on MS-COCO captions, marking specific object mentions as either hallucinated or faithful. It provides two key metrics *CHAIRs*, the percentage of generated captions containing at least one hallucinated object, and *CHAIRi*, the percentage of hallucinated object instances among all object mentions in the generated captions. Following previous papers, we randomly selected 500 samples from MS-COCO validation set for our experiments.

Model	CHAIR$_S$ ↓	CHAIR$_I$ ↓	Recall ↑
LLaVA-1.5	51.0 ↑0.0	14.7 ↑0.0	82.8 ↑0.0
+ VCD	47.8 ↓3.2	14.1 ↓0.6	82.7 ↓0.1
+ VDD-None	50.2 ↓0.8	14.3 ↓0.4	**83.2** **↑0.4**
+ ASD (Ours)	**40.0** **↓11.0**	**11.3** **↓3.4**	82.0 ↓0.8

Table 2: Comparison of different hallucination mitigation methods on CHAIR benchmark. CHAIR$_S$ and CHAIR$_I$ measure sentence-level and instance-level hallucination rates respectively (lower is better), while Recall measures the model's ability to describe actually present objects (higher is better). Our method achieves substantial reductions in hallucination rates with only minimal impact on recall performance.

General Visual Understanding. **MME** is a comprehensive benchmark designed to assess LVLMs through yes/no questions. It comprises 14 subsets: 10 perception-based tasks (including color, count, position, scene, action, etc.) and 4 reasoning-based tasks (including commonsense, numerical, mathematical reasoning). **MMBench** is a comprehensive multiple-choice benchmark containing approximately 3,000 questions across 20 ability dimensions covering perception and reasoning tasks. We use the DEV split containing 1,164 English questions for evaluation. **MMMU** is a challenging multiple-choice benchmark containing 11.5K questions spanning 30 academic subjects at the college level. The benchmark is particularly challenging, with even GPT-4V achieving less than 60% accuracy. **TextVQA** validation set consists of 5,000 questions that can only be correctly answered by reading and reasoning about text present in images. **LLaVA-Bench** consists of 60 carefully designed open-ended questions across 24 images, evaluating models' visual reasoning and understanding capabilities. The responses are evaluated using GPT-4-1106-preview as an automatic evaluator, providing standardized scoring metrics. **MM-Vet** contains 217 challenging open-ended tasks that require models to simultaneously demonstrate multiple capabilities including detailed perception, cross-modal reasoning, and world knowledge. We use the official online evaluator, powered by GPT-4-0613, to ensure fair comparison with existing approaches.

6.2 Implementation Details

We conduct experiments on two base model: LLaVA1.5-7B (Liu et al., 2024b) and Qwen-VL-Chat (Bai et al., 2023). For each model, we randomly sample 1,000 images from MSCOCO training set for steering vector extraction of Eq. (1). We conduct grid search over $\lambda \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ for both π^+ and π^-. For comparison, we implement VCD (Leng et al., 2024) with optimized hyperparameters, and VDD-None (Zhang et al., 2024) using their recommended parameters.

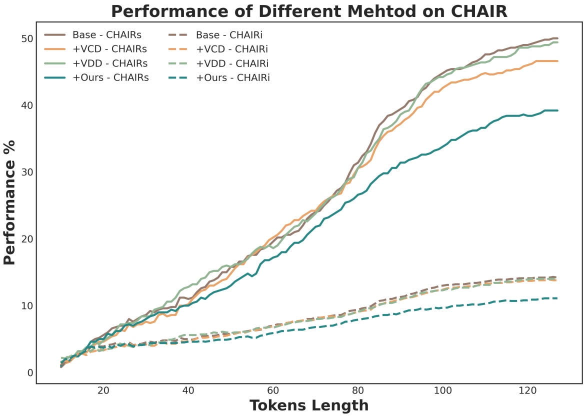
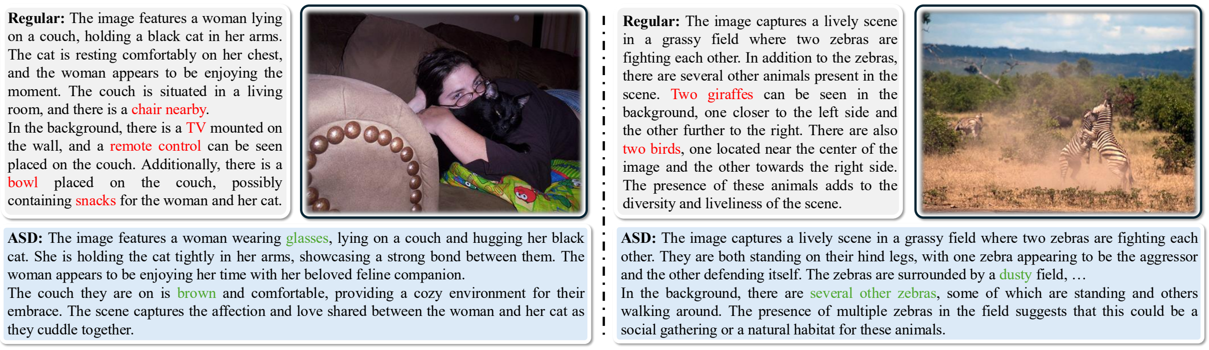


Figure 3: Analysis of hallucination rates (CHAIR$_S$ and CHAIR$_I$) with respect to generated token length, with LLaVA1.5-7b as the base model.

6.3 Hallucination Reduction Performance

Tab. 1 presents a comprehensive evaluation of our method against existing approaches on the POPE benchmark. We evaluate performance under two decoding strategies: Greedy Decoding and Direct Sampling (which generates responses by directly sampling from the raw logit probability distribution without normalization) across three subset (MSCOCO, A-OKVQA, and GQA), using both accuracy and F1 score as metrics. Our method demonstrates consistent and substantial improvements across all experimental settings. Under Greedy Decoding, when applied to LLaVA1.5-7B, our approach achieves absolute gains of 2.88%, 6.11%, and 6.89% in accuracy on MSCOCO, A-OKVQA, and GQA respectively. The improvements were even more pronounced when applied to Qwen-VL-Chat, particularly on the GQA dataset where we observed a remarkable **8.54%** increase in accuracy and **14.51%** improvement in F1 score. Notably, our method not only surpasses the baseline models but also outperforms existing hallucination mitigation approaches (VCD and VDD-None) by a significant margin. The effectiveness of our method is further validated under Direct Sampling, where it maintains robust performance improvements. For instance, with LLaVA1.5-7B, our method achieves accuracy gains of 5.70%, 8.66%, and 9.48% on the three subset respectively. Unlike other methods showing more significant improvements under direct sampling, our approach demonstrates robust



Regular: The image features a woman lying on a couch, holding a black cat in her arms. The cat is resting comfortably on her chest, and the woman appears to be enjoying the moment. The couch is situated in a living room, and there is a chair nearby.
In the background, there is a TV mounted on the wall, and a remote control can be seen placed on the couch. Additionally, there is a bowl placed on the couch, possibly containing snacks for the woman and her cat.

ASD: The image features a woman wearing glasses, lying on a couch and hugging her black cat. She is holding the cat tightly in her arms, showcasing a strong bond between them. The woman appears to be enjoying her time with her beloved feline companion.
The couch they are on is brown and comfortable, providing a cozy environment for their embrace. The scene captures the affection and love shared between the woman and her cat as they cuddle together.

Regular: The image captures a lively scene in a grassy field where two zebras are fighting each other. In addition to the zebras, there are several other animals present in the scene. Two giraffes can be seen in the background, one closer to the left side and the other further to the right. There are also two birds, one located near the center of the image and the other towards the right side. The presence of these animals adds to the diversity and liveliness of the scene.

ASD: The image captures a lively scene in a grassy field where two zebras are fighting each other. They are both standing on their hind legs, with one zebra appearing to be the aggressor and the other defending itself. The zebras are surrounded by a dusty field, …
In the background, there are several other zebras, some of which are standing and others walking around. The presence of multiple zebras in the field suggests that this could be a social gathering or a natural habitat for these animals.

Figure 4: Illustration of ASD correcting hallucinations. Hallucinated objects (red) are removed while ASD adds accurate visual details (green).

effectiveness under both greedy decoding and direct sampling strategies, validating its stability and reliability across different inference settings. The superior performance can be attributed to our contrast decoding mechanism, which effectively isolates and suppresses hallucination patterns while preserving the model's ability to generate accurate and contextually appropriate responses. This is evidenced by the consistent improvements across both metrics and all datasets, suggesting that our method successfully addresses hallucination without compromising general visual understanding capabilities.

The result on the CHAIR benchmark is reported in Tab. 2. Our method demonstrates substantial improvements in reducing hallucination rates compared to the baseline LLaVA1.5-7B model and other mitigation approaches. Specifically, we achieve a significant 10.0% reduction in sentence-level hallucination (CHAIR$_S$) compared to the baseline, substantially outperforming both VCD (-3.2%) and VDD-None (-0.8%). The CHAIR$_I$ metric exhibited a similar trend. Notably, while VDD-None achieves the best recall performance with a 0.4% improvement over the baseline, our method still maintains competitive recall (-0.8%) while achieving significantly better hallucination reduction, demonstrating a favorable trade-off between reliability and comprehensiveness. This minimal trade-off in recall suggests that our approach effectively reduces hallucination while largely preserving the model's ability to describe actually present objects in the images.

Fig. 3 illustrates the relationship between generated token length and hallucination rates across different methods, where the base model is LLaVA1.5-7B. Our analysis reveals that hallucination rates increase progressively with the length of generated content across all methods. A particularly concerning observation is the presence of a sharp increase in hallucination rates around the 80-token mark across all methods, suggesting that extended generation lengths pose heightened risks for hallucination. Notably, our approach demonstrates particularly strong advantages beyond this threshold, maintaining substantially lower hallucination rates with a notably smaller slope in both CHAIR$_S$ and CHAIR$_I$ metrics compared to baseline and existing methods.

Visualization example. To provide concrete illustrations of how ASD mitigates hallucinations in practice, we present qualitative comparisons in Fig. 4. In the left example, the baseline hallucinates multiple objects (chair, TV, remote control, bowl, snacks) that are absent from the image, while ASD removes these errors and accurately describes the woman with her cat, adding new details like "wearing glasses" and "brown couch." Similarly, in the right example, ASD corrects the baseline's hallucinations of "giraffes" and "birds," properly describing only the zebras while introducing additional accurate details about the "dusty field" environment. These results illustrate how ASD not only suppresses hallucination patterns but also enhances descriptive richness with factually accurate details.

6.4 General Performance Maintenance

Tab. 3 presents the results on six general visual understanding benchmarks. Our method shows comparable or improved performance across most tasks for both models. For LLaVA1.5-7B, we observe notable improvements on MME (+16.51), MMMU (+3.34), and MMVet (+2.70) while maintaining performance on other benchmarks with minimal variation. Similarly, for Qwen-VL-Chat, our method achieves the best performance on MMMU (+3.00), MMVet (+0.50), and LLaVABench (+1.80), with negligible degradation on other benchmarks. This

Method	**MME**	**MMBench**	**MMMU**	**TextVQA**	**MMVet**	**LLaVABench**	**Overall**
LLaVA1.5-7B	1810.70 $\uparrow$0.00	**65.46** **$\uparrow$0.00**	35.44 $\uparrow$0.00	45.76 $\uparrow$0.00	31.10 $\uparrow$0.00	58.90 $\uparrow$0.00	$\uparrow$0.00
+ VCD	1800.41 $\downarrow$10.29	64.69 $\downarrow$0.77	36.00 $\uparrow$0.56	44.26 $\downarrow$1.50	30.90 $\downarrow$0.20	57.20 $\downarrow$1.70	$\downarrow$4.18
+ VDD-None	1763.80 $\downarrow$46.90	63.75 $\downarrow$1.71	36.78 $\uparrow$1.34	42.19 $\downarrow$3.57	32.30 $\uparrow$1.20	**62.10** **$\uparrow$3.20**	$\downarrow$2.13
+ ASD (Ours)	**1827.21** **$\uparrow$16.51**	65.38 $\downarrow$0.08	**38.78** **$\uparrow$3.34**	**46.40** **$\uparrow$0.64**	**33.80** **$\uparrow$2.70**	61.60 $\uparrow$2.70	**$\uparrow$10.21**
Qwen-VL-Chat	1839.55 $\uparrow$0.00	61.34 $\uparrow$0.00	33.56 $\uparrow$0.00	**60.79** **$\uparrow$0.00**	46.10 $\uparrow$0.00	66.40 $\uparrow$0.00	$\uparrow$0.00
+ VCD	1847.85 $\uparrow$8.30	60.40 $\downarrow$0.94	35.67 $\uparrow$2.11	59.31 $\downarrow$1.48	45.20 $\downarrow$0.90	67.50 $\uparrow$1.10	$\uparrow$0.34
+ VDD-None	**1861.01** **$\uparrow$21.46**	**62.97** **$\uparrow$1.63**	33.67 $\uparrow$0.11	59.91 $\downarrow$0.88	41.40 $\downarrow$4.70	65.20 $\downarrow$1.20	$\downarrow$3.87
+ ASD (Ours)	1825.20 $\downarrow$14.35	61.08 $\downarrow$0.26	**36.56** **$\uparrow$3.00**	60.42 $\downarrow$0.37	**46.60** **$\uparrow$0.50**	**68.20** **$\uparrow$1.80**	**$\uparrow$3.89**

Table 3: Performance comparison on general visual understanding benchmarks. Bold numbers indicate the best scores for each benchmark. When calculating overall improvements, percentage changes are used for MME scores and absolute changes for other benchmarks due to scale differences. Results show that our method maintains or improves performance across diverse tasks compared to baseline models and other approaches.

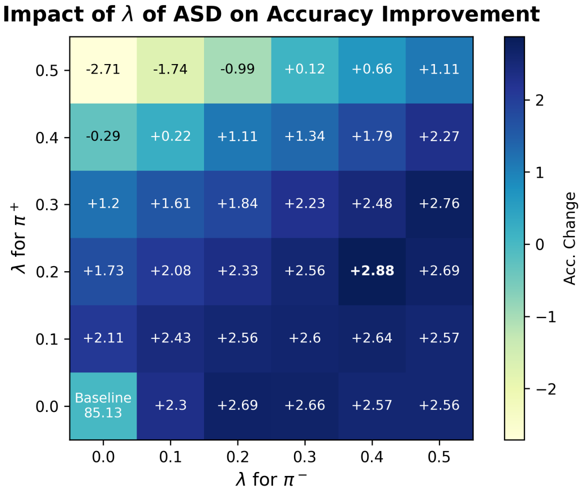


Figure 5: Impact of steering intensities on ASD, measured as percentage point improvements over LLaVA1.5-7B baseline (85.13%) on POPE-COCO accuracy. The optimal performance (+2.88%) is achieved with $\lambda = 0.2$ for π^+ and $\lambda = 0.4$ for π^-.

Count	**POPE**	**CHAIR$_S$ $\downarrow$**	**MME**	**TextVQA**
LLaVA1.5-7B	85.13	51.00	1810.70	45.76
100	87.72	40.40	1813.01	46.22
500	87.79	**38.80**	1821.98	46.24
1,000	**88.01**	40.00	**1827.21**	**46.40**

Table 4: Impact of calibration data size (number of images used for steering vector computation) on model performance across different benchmarks. POPE refers to POPE-COCO subset.

dual achievement - substantial hallucination reduction while preserving and sometimes improving general capabilities - validates the effectiveness of our contrast decoding mechanism in mitigating hallucination patterns without compromising essential visual understanding features.

6.5 Ablation Study

6.5.1 Impact of Steering Strength

Fig. 5 illustrates the effect of steering intensities λ of ASD method. Most parameter combinations yield positive improvements over the baseline, demonstrating the robustness of our method. However, we observe that positive steering (π^+) requires more careful tuning - performance begins to degrade when $\lambda > 0.3$, with accuracy dropping by 2.71% at $\lambda = 0.5$. In contrast, negative steering (π^-) shows greater tolerance to larger values, maintaining improvements even at $\lambda = 0.5$. The optimal configuration is achieved with moderate positive steering ($\lambda = 0.2$ for π^+) and stronger negative steering ($\lambda = 0.4$ for π^-), achieving **88.01%** accuracy (a 2.88% improvement over the baseline), which represents a state-of-the-art performance on this benchmark.

6.5.2 Impact of Calibration Data Size

Tab. 4 examines the sensitivity of our method to the amount of calibration data used for computing steering vectors. Notably, our approach demonstrates strong performance even with mini calibration data - using just **only 100** images already yields substantial improvements across all selected benchmarks. These results suggest that our method can effectively capture hallucination patterns with a very small calibration set, making it highly practical for real-world applications.

6.5.3 Direct Vector Steering

We investigate the effectiveness of vector steering without contrast decoding to understand its impact in isolation. Fig. 6 shows the accuracy improvements over the LLaVA1.5-7B baseline on POPE benchmark. The y-axis represents the relative accuracy change in percentage points compared to the baseline performance. First, we observe that the optimal steering intensity varies significantly across datasets, with COCO achieving peak performance at $\lambda = 0.3$, while AOKVQA and GQA show improvements at lower intensities. This variation

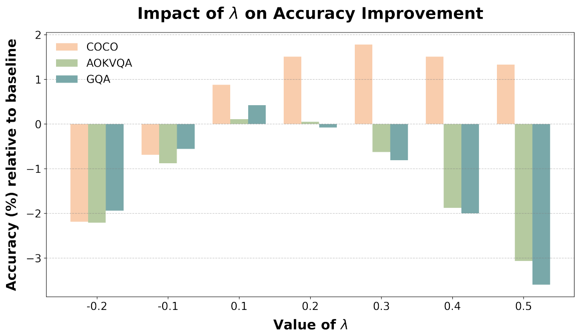


Figure 6: Impact of steering intensity on Direct Vector Steering, measured as relative improvement over LLaVA1.5-7B baseline.

suggests that the effectiveness of steering vectors is sensitive to the specific characteristics of each task. Second, we observe a consistent pattern where performance deteriorates at higher steering intensities. This degradation becomes particularly pronounced at $\lambda = 0.5$, where AOKVQA and GQA show accuracy drops of approximately 3% and 3.5% respectively. This decline can be attributed to excessive distortion of the hidden state semantics, indicating that overly aggressive steering can disrupt the model's learned representations. While COCO shows substantial improvements of up to 1.8%, the gains on AOKVQA and GQA are notably smaller. This performance gap is expected, as the calculation of steering vectors relies on COCO-defined object categories. This suggests that direct vector steering may have limitations in generalizing across different visual understanding tasks.

6.6 Computational Efficiency Analysis

While ASD requires performing addition and subtraction operations on hidden states at each layer during inference with two branches, the computational overhead is relatively modest for several practical reasons.

First, the addition and subtraction operations on hidden states are extremely lightweight compared to the model's transformer operations (self-attention and feed-forward computations).

Second, although our method involves two branches for positive and negative steering, we leverage batch processing to run them in parallel during inference. Specifically, we concatenate the inputs for both positive and negative steering along the batch dimension, allowing them to be processed simultaneously. This parallel processing approach significantly mitigates the potential overhead compared to sequential execution.

To quantify the exact overhead, we conduct timing experiments on LLaVA-1.5-7B using an NVIDIA L20 GPU. As shown in Tab. 5, with parallel processing strategy, our method introduces approximately 10-20% additional computational overhead. This characteristic is particularly beneficial for real-world applications on edge devices, where the query is typically limited to 1, and GPUs often have spare capacity to handle a batch size of 2. This makes our method an efficient solution that offers significant accuracy improvements with minimal computational cost.

Method	20 Tokens (ms)	50 Tokens (ms)	100 Tokens (ms)
LLaVA-1.5-7B	516.9 ±8.1	1153.1 ±7.4	2211.7 ±6.9
+ ASD (Ours)	612.8 ±7.2	1300.2 ±7.2	2467.8 ±8.1
Overhead (%)	18.6%	12.8%	11.6%

Table 5: Inference time (in milliseconds) comparison between the baseline and the ASD method across different output lengths. The ASD method introduces moderate additional latency, which becomes relatively smaller as the number of generated tokens increases.

7 Conclusion

We present a systematic investigation of hallucination in LVLMs through the lens of intermediate activations, revealing that hallucinated content manifests as distinct patterns in the model's hidden state space. Building on this insight, we propose Activation Steering Decoding, a training-free approach that effectively mitigates hallucination through targeted intervention in model activations. Our extensive experiments demonstrate that our approach significantly reduces hallucination rates while maintaining model performance across general visual understanding tasks.

Limitations

While our proposed Activation Steering Decoding demonstrates promising results in mitigating hallucination, several limitations warrant discussion.

First, our current approach primarily addresses object-level hallucination, as the steering vectors are extracted using only COCO object annotations. This focus on object categories limits the method's ability to address other types of hallucinations, such as attribute errors (e.g., incorrect colors or sizes), relational inaccuracies (e.g., wrong spatial relationships), or hallucinations involving abstract concepts and actions. Future work should explore leveraging richer annotations beyond object labels to develop more comprehensive hallucination mitigation strategies.

Second, our approach involves hyperparameters (λ for steering intensity and α for contrast coefficient), which may vary across different models and tasks for optimal performance. Developing adaptive approaches that can automatically determine optimal steering parameters based on the input or model confidence remains an interesting direction for future research.

References

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Guang Dai, Ping Chen, and Shijian Lu. 2024. Agla: Mitigating object hallucinations in large vision-language models with assembly of global and local attention. *arXiv preprint arXiv:2406.12718*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. 2023. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*.

Ailin Deng, Zhirui Chen, and Bryan Hooi. 2024. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*.

Xin He, Longhui Wei, Lingxi Xie, and Qi Tian. 2024. Incorporating visual experts to resolve the information loss in multimodal large language models. *arXiv preprint arXiv:2401.03105*.

Jitesh Jain, Jianwei Yang, and Humphrey Shi. 2024. Vcoder: Versatile vision encoders for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27992–28002.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

Qirui Jiao, Daoyuan Chen, Yilun Huang, Yaliang Li, and Ying Shen. 2024. Enhancing multimodal large language models with vision detection models: An empirical study. *arXiv preprint arXiv:2401.17981*.

Ole Jorgensen, Dylan Cope, Nandi Schoots, and Murray Shanahan. 2023. Improving activation steering in language models with mean-centring. *arXiv preprint arXiv:2312.03813*.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2023. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. 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*, pages 13872–13882.

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

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Wenhao Liu, Xiaohua Wang, Muling Wu, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. 2023b. Aligning large language models with human preferences through representation engineering. *arXiv preprint arXiv:2312.15997*.

Varun Magesh, Faiz Surani, Matthew Dahl, Mirac Suzgun, Christopher D Manning, and Daniel E Ho. 2024. Hallucination-free? assessing the reliability of leading ai legal research tools. *arXiv preprint arXiv:2405.20362*.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2023. Steering llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681*.

Nishant Subramani, Nivedita Suresh, and Matthew E Peters. 2022. Extracting latent steering vectors from pretrained language models. *arXiv preprint arXiv:2205.05124*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *arXiv e-prints*, pages arXiv–2308.

Jue Wang. 2024. Hallucination reduction and optimization for large language model-based autonomous driving. *Symmetry*, 16(9):1196.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2024. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pages 32–45. Springer.

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

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2024a. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953.

Runpeng Yu, Weihao Yu, and Xinchao Wang. 2025. Attention prompting on image for large vision-language models. In *European Conference on Computer Vision*, pages 251–268. Springer.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024b. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Zihao Yue, Liang Zhang, and Qin Jin. 2024. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *arXiv preprint arXiv:2402.14545*.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2023. Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *arXiv preprint arXiv:2310.01779*.

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

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. 2024. Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv preprint arXiv:2402.08680*.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*.

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

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2024. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.