SHARP: Steering Hallucination in LVLMs via Representation Engineering

Junfei Wu[1,2*], **Ding Yue**[1,2,*], **Guofan Liu**[1,2], **Tianze Xia**[2], **Ziyue Huang**[2],
Dianbo Sui[3], **Qiang Liu**[1,2,†], **Shu Wu**[1,2], **Liang Wang**[1,2], **Tieniu Tan**[1,2,4]
[1] New Laboratory of Pattern Recognition (NLPR),
State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, Chinese Academy of Sciences
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Harbin Institute of Technology [4]Nanjing University
junfei.wu@cripac.ia.ac.cn, {yue.ding, qiang.liu}@nlpr.ia.ac.cn

Abstract

Despite their impressive capabilities, Large Vision-Language Models (LVLMs) frequently generate plausible yet incorrect or unsupported responses, referred to as hallucinations. In this study, we investigate whether different types of hallucinations are reflected in the model's internal representations by probing their encoded features. We focus on two causes of hallucination in multimodal reasoning—(1) over-reliance on textual priors and (2) preference for user prompts over conflicting visual evidence—which have been identified in prior work as frequent and impactful factors. Our probing results reveals that hallucinations exhibit distinguishable representational patterns, suggesting a representation-level approach to characterize and mitigate them. Motivated by this, we propose **S**teering **HA**llucination via **R**e**P**resentation Engineering (**SHARP**), a representation-level intervention framework that modulates hallucination-related features during inference. SHARP identifies functional representations responsible for prior-driven and visual-context conflicts, and jointly adjusts the model's internal activations during inference. We evaluate our approach extensively using three large vision-language models across various benchmarks. Experimental results show that our proposed intervention effectively reduces hallucinations without compromising the performance and generalization of the LVLMs.

1 Introduction

Large Vision-Language Models (LVLMs) have demonstrated exceptional capabilities across a wide range of multimodal tasks (Liu et al., 2023; Bai et al., 2023; Hurst et al., 2024; Jaech et al., 2024), spanning from basic perception to recognition and complex reasoning. However, they are inevitably plagued by hallucination issues-generating content that contradicts the given multimodal context. This limitation not only hinders model reliability but also poses serious safety concerns (Liu et al., 2024b). Identifying the underlying mechanisms responsible for hallucinations can enhance our understanding of LVLMs' limitations and pave the way for developing more reliable models.

There have been several studies that investigate the causes of hallucinations in LVLMs from two perspectives: external and internal. From the external perspective, hallucinations are often attributed to data-level biases. For instance, POPE (Li et al., 2023) and LURE (Zhou et al., 2024) show that hallucinated objects are often popular ones in training data or those that co-occur with objects mentioned in the instructions or prior responses. Additionally, LRV-Instruction (Liu et al., 2024a) highlights that most LVLMs are fine-tuned to encourage them to cater to positive instructions.

In contrast, internal-factor analyses attribute hallucination to architectural behaviors. Contrastive decoding-based methods, such as VCD (Leng et al., 2024) and HALC (Chen et al., 2024b), share the common assumption that hallucinations arise from the model's sensitivity to image and instruction inputs, coupled with unimodal biases inherited from LLMs. Additionally, studies like Opera (Huang et al., 2024) and AD-HH/TF-HH (Yang et al., 2024) reveal that some specific attention patterns within attention heads can cause LVLMs to overlook critical visual information, ultimately leading to hallucinated outputs.

Despite these advancements, existing methods primarily associate overall hallucination patterns with internal model behaviors (Huang et al., 2024; Yang et al., 2024), without disentangling the causes of different types. For instance, it remains unclear whether the model contains internal features that can distinguish between hallucinations driven by textual priors and those induced by vision-text conflicts. This gap hinders the development of targeted interventions that can precisely address distinct hal-

[*]Equal Contribution. [†]Corresponding Author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 14358–14373
November 4-9, 2025

lucination causes.

To fill this gap, we examine two causes of hallucination in multimodal reasoning—over-reliance on textual priors and vision-context conflicts—which have been identified in prior work as frequent and impactful causes of hallucination (Bitton-Guetta et al., 2023; Wang et al., 2024; Liu et al., 2024f; Bai et al., 2024; Leng et al., 2024). Our analysis spans intra- and inter-cause levels: the former examines whether internal representations differ under a single cause, while the latter investigates whether there are distinguishable features associated with hallucinations arising from different causes. As shown in Fig. 1, the model not only detects cause-specific hallucinations but also differentiates between different hallucination categories, revealing an inherent capacity for fine-grained, representation-level intervention. The detailed analysis is listed in Sec. 3.

Based these insights, we propose SHARP (**S**teering **HA**llucination via **R**e**P**resentation Engineering), a novel inference-time intervention strategy designed to mitigate hallucinations in Large Vision-Language Models (LVLMs). SHARP leverages decomposed steering vectors to modulate internal representations and reduce hallucination without retraining. As illustrated in Fig. 2, our method consists of three key stages: (1) Stimulus-driven data collection, where we elicit model responses under different hallucination-inducing conditions; (2) Cause-specific vector derivation, in which we extract steering vectors by contrasting hidden activations between faithful and hallucinated responses for each casues; (3) Hallucination intervention: different steering vectors are jointly applied during inference to mitigate hallucination-inducing patterns. Comprehensive experimental results show that SHARP achieves a significant improvement in hallucination reduction compared to existing methods.

Our contributions are summarized as follows:

- We demonstrate that LVLMs' internal representations encode informative signals associated with distinct hallucination causes, indicating the presence of internal cues that reflect their underlying causes even when hallucinations occur.
- We propose SHARP, a novel inference-time method that steers cause-specific activations to mitigate hallucinations by adjusting internal states.
- Extensive experiments on multiple benchmarks demonstrate that SHARP significantly reduces hallucinations while preserving generation capabilities.

2 Related Work

2.1 Hallucination Mitigation Methods in LVLMs

Hallucination has been a critical challenge in LVLMs (Li et al., 2023; Liu et al., 2024c). Unlike hallucination mitigation in LLMs (Manakul et al.; Chuang et al., 2024; Zhang et al., 2025), which primarily targets improving factual consistency in text generation, mitigation in LVLMs aims to align model outputs with visual evidence. A range of methods have been proposed for this purpose, which can be broadly grouped into training-based approaches (Sun et al., 2024; Gunjal et al., 2024; Jiang et al., 2024), inference-time interventions (Leng et al., 2024; Chen et al., 2025; Wang et al., 2025; Kim et al., 2024), and post-generation correction methods (Zhou et al., 2024; Yin et al., 2024; Chen et al., 2024a).

Training-based approaches, such as LRV-Instruction (Liu et al., 2024a) and HACL (Jiang et al., 2024), improve LVLM robustness by constructing diverse and comprehensive datasets for instruction tuning. Inference-time strategy have demonstrated effectiveness and efficiency in reducing hallucinations. For instance, VCD (Leng et al., 2024) compares outputs under perturbed visual inputs to suppress over-reliance on unimodal priors. OPERA (Huang et al., 2024) detects abnormal attention patterns and applies a rollback mechanism to penalize hallucination-prone behaviors. HALC (Chen et al., 2024b) further integrates external grounding signals to enforce both local and global visual-textual consistency during decoding. In addition, CLIP-guided scoring (Deng et al., 2024) ranks candidate responses by evaluating their alignment with visual input. Post-generation correction methods, such as LURE (Zhou et al., 2024) and LogicCheckGPT (Wu et al., 2024), refine model outputs by detecting and revising hallucinations through external verification or internal consistency checks.

Compared to prior methods, our approach directly intervenes in the latent representation space of LVLMs to suppress hallucination behavior. It performs a single-step representation-level modification during inference, achieving both effec-

tiveness and efficiency without retraining or extra overhead.

2.2 Representation Engineering Methods

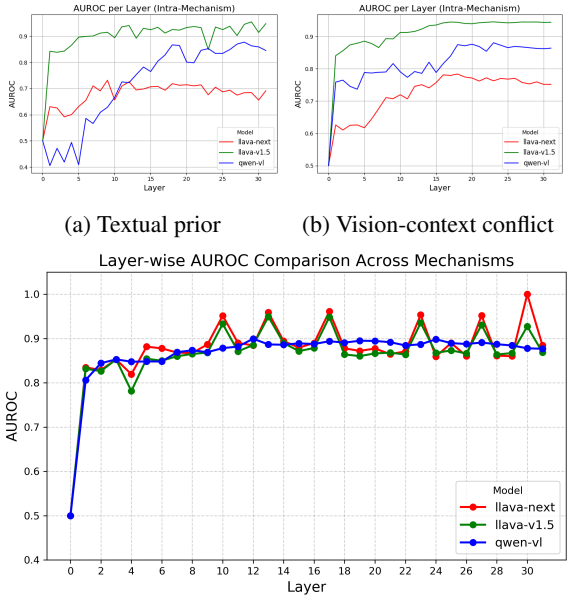
Recent efforts to interpret and steer LLMs have increasingly turned to leverage representation engineering, which manipulates internal representations without full retraining. This approach enables scalable alignment of model behavior. For example, ActAdd (Turner et al., 2023) identified directional vectors corresponding to specific concepts, while RepE (Zou et al., 2023) showed that vector operations in activation space can steer factuality in generation. Beyond intervention, some methods extract latent vectors representing abstract features and leverage them to guide inference, producing more aligned outputs through controlled prompting or gating mechanisms (Subramani et al., 2022; Panickssery et al., 2023).

Inspired by these works, we extend representation engineering to LVLMs. By identifying hallucination-related directions and applying lightweight, training-free interventions at specific layers, we effectively suppress hallucinations in LVLMs without sacrificing performance.

3 Analysis: Diagnosing Hallucination via Representation Separability

We investigate two causes of hallucination in LVLMs, both of which have been recognized as frequent and significant contributors in prior studies: (1) **Textual priors**, where the model overrelies on patterns learned during language pretraining, generating answers consistent with linguistic co-occurrence or commonsense associations even when they contradict the visual input. (2) **Vision-context conflicts**, where misleading or false assumptions in the prompt conflict with the visual content, testing whether the model can rely on visual evidence rather than textual cues. These two causes represent distinct failure modes in LVLMs, illustrating different pathways through which hallucinations can occur.

To test whether internal representations encode hallucination signals, we construct cause-specific datasets $\mathcal{D}^{(m)} = \mathcal{D}^{(m)}_{\text{faithful}} \cup \mathcal{D}^{(m)}_{\text{hallucinated}}$ for each cause $m \in \{\text{T}, \text{C}\}$, where T denotes textual priors and C denotes vision–context conflicts. Each contains query-answer pairs labeled by whether the response is faithful to the visual input, enabling intra-cause analysis. We also define an inter-cause dataset $\mathcal{D}_{\text{inter}} = \mathcal{D}^{(\text{T})} \cup \mathcal{D}^{(\text{C})}$ to examine shared and distinct features across the two hallucination types. The details about data construction can be referred in 4.1.



(a) Textual prior (b) Vision-context conflict

(c) Comparison across causes

Figure 1: Probing results across layers for LLaVA-v1.5-7B, Qwen-VL, and LLaVA-Next. (a) and (b) show intra-cause probing, where a logistic regression classifier is trained to distinguish hallucinated from non-hallucinated samples under each specific hallucination cause (e.g., textual prior or vision-context conflict). (c) shows inter-cause probing, where a logistic regression classifier is trained to differentiate hallucinated samples originating from different hallucination causes. Models exhibit clear separability in both settings.

We analyze the residual stream at the final input position, which encodes both image and question context and predicts the first response token, to probe for hallucination signals. We employ linear probing to test whether residual activations encode hallucination signals. Logistic regression classifiers are trained to distinguish between faithful and hallucinated responses (intra-cause), and between hallucination types (inter-cause), using activations from each layer. Evaluation on held-out data with AUROC scores is conducted for both LLaVA-v1.5-7B, LLaVA-Next and Qwen-VL.

The probing results are shown in Fig. 1. Specifically, Fig. 1a and Fig. 1b demonstrate as layer depth increases, the separation between hallucinated and factual examples under textual prior and vision-context conflict causes becomes clearer and stabilizes around layers 10 to 15. This suggests that the models progressively encode discriminative fea-

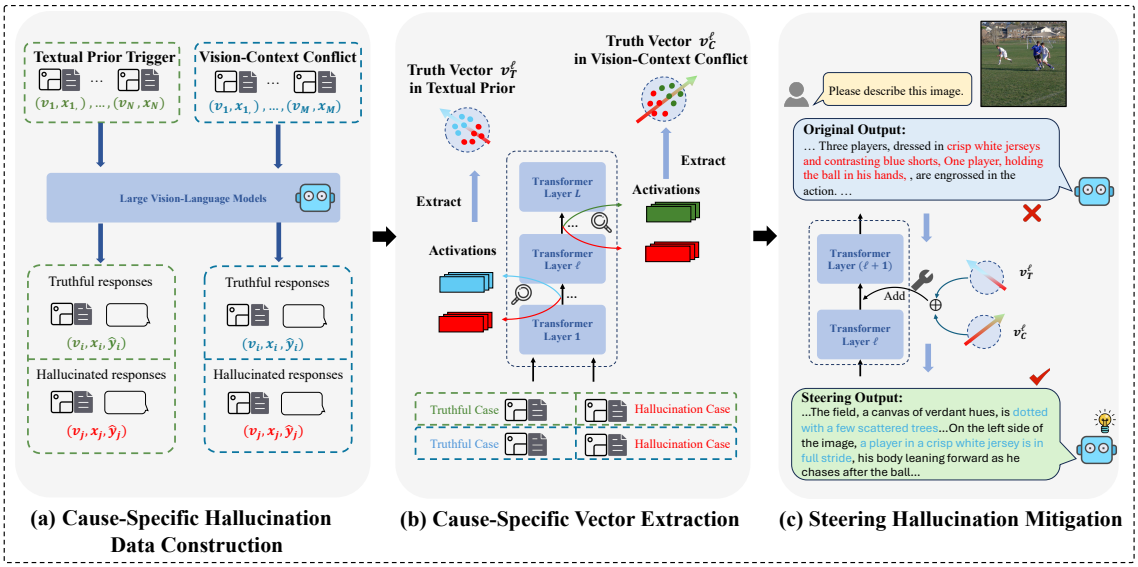


Figure 2: The overall framework of SHARP for steering hallucination behavior.

tures tied to each hallucination type, with early to mid layers playing a key role in developing feature separability. Fig. 1c further shows that different hallucination types are also distinguishable from one another, suggesting that the models implicitly capture the nature of the hallucination trigger. These findings demonstrate LVLMs' internal sensitivity to hallucination cues and support our training-free, inference-time mitigation strategy leveraging this latent signal.

4 Method

Our method, SHARP, aims to identify two functional vectors that are closely related to hallucination causes in LVLMs, using contrastive activation analysis. These vectors are then used to guide model generation through intervention on internal activations. As illustrated in Fig. 2, SHARP operates in three stages: (1) Constructing cause-specific data to stimulate hallucination-related activations; (2) Extracting cause-relevant direction vectors via contrastive analysis; and (3) Steering model activations during inference based on these vectors.

4.1 Cause-Stimulated Data Construction

To systematically analyze the underlying causes of hallucinations in MLLMs, we construct a cause-oriented dataset $\mathcal{D}$, which is divided into two subsets based on the types of hallucination-inducing causes: (1) the over-reliance on textual priors subset $\mathcal{D}^{\mathrm{T}}$, where the question is designed to induce hallucinations by exploiting linguistic priors, and (2) the visual-textual conflict subset $\mathcal{D}^{\mathrm{C}}$, where the textual query contradicts the visual content of the image. These two causes are empirically associated with hallucination generation in LVLMs. Specifically, we adopt the *insufficient context* subset from the HaloQuest benchmark (Wang et al., 2024) as $\mathcal{D}^{\mathrm{T}}$, and the *false premises* subset as $\mathcal{D}^{\mathrm{C}}$. Detailed information about the benchmark is provided in AppendixB.1.

Given an image-question pair (v_i, x_i) from the dataset, where v_i is the input image and x_i is the corresponding natural language question, we query a base multimodal model M (e.g., LLaVA or LLaVA-Next) to obtain its predicted answer:

$$\hat{y}_i = M(v_i, x_i) \tag{1}$$

Then we leverage the LLM-as-a-Judge framework to assess the factual correctness of the generated answer $\hat{y}_i$. Specifically, we provide the scoring model (GPT-4o-mini) with both the model-generated answer $\hat{y}_i$ and the reference ground truth answer y_i for each input pair (v_i, x_i). The scoring model returns a binary factuality label:

$$s_i = \mathrm{Judge}(y_i, \hat{y}_i) \in \{\mathrm{Correct}, \mathrm{Incorrect}\} \tag{2}$$

where `Correct` indicates that the generated answer is factually consistent with the ground truth. The exact prompt design used in the scoring process is provided in AppendixB.1.

Based on the factuality label s_i for each multimodal input, we further divide the cause-specific

subsets into factual and hallucinatory splits. Specifically, $\mathcal{D}^{\text{T}}$ is split into:

$$\mathcal{D}^{\text{T}} = \mathcal{D}^{\text{T}}_{\text{truth}} \cup \mathcal{D}^{\text{T}}_{\text{hallucination}} \tag{3}$$

where $\mathcal{D}^{\text{T}}_{\text{truth}}$ contains examples where the model's answers are correct, and $\mathcal{D}^{\text{T}}_{\text{hallucination}}$ includes those where hallucinations occur. Similarly, we divide $\mathcal{D}^{\text{C}}$ into:

$$\mathcal{D}^{\text{C}} = \mathcal{D}^{\text{C}}_{\text{truth}} \cup \mathcal{D}^{\text{C}}_{\text{hallucination}} \tag{4}$$

These four refined subsets represent multimodal inputs that elicit different behaviors from the model when confronted with distinct hallucination-inducing factors. Analyzing these subsets enables us to probe the internal representations that help the model suppress hallucinations and enhance factual alignment under different cause conditions.

4.2 Cause Vector Extraction

We extract cause-specific steering vectors by performing contrastive analysis over the model's hidden representations. Let $h^{\ell}(v, x) \in \mathbb{R}^d$ denote the hidden state of the final token at layer ℓ when the model processes an input image v and question x.

To derive a steering vector for each hallucination cause, we aggregate contrastive signals across all samples in the corresponding cause-specific subset. For a cause type m, given the dataset $\mathcal{D}^{(m)} = \{(v_i, x_i, \hat{y}_i)\}_{i=1}^{N}$, we compute:

$$\vec{v}^{\ell}_{\text{m}} = \frac{1}{|\mathcal{D}^{(m)}_{\text{truth}}|} \sum_{(v_i,x_i)\in\mathcal{D}^{(m)}_{\text{truth}}} h^{\ell}(v_i, x_i) - \frac{1}{|\mathcal{D}^{(m)}_{\text{hallucination}}|} \sum_{(v_j,x_j)\in\mathcal{D}^{(m)}_{\text{hallucination}}} h^{\ell}(v_j, x_j) \tag{5}$$

In practice, we normalize $\vec{v}^{\ell}_{\text{m}}$ to ensure consistent scaling across layers and cause types. This steering vector captures the average discriminative direction in the representation space that distinguishes truthful from hallucinated responses under cause m. These cause-specific vectors serve as interpretable directions in the activation space and they can be used to probe the model's reasoning or directly intervene in hidden states, enabling us to steer model toward more faithful outputs.

4.3 Steering Hallucination Mitigation

After obtaining the cause-specific steering vectors (Section 4.2), we aim to steer the model by directly intervening in its internal representations during inference. Specifically, we select a target layer ℓ^* and inject a linear combination of $\vec{v}^{\ell^*}_{\text{T}}$ and $\vec{v}^{\ell^*}_{\text{C}}$ into the hidden states of each generated token $t \geq |x_i|$:

$$h^{(\ell^*)}_{\text{steered}}(v_i, x_i)_t = h^{(\ell^*)}(v_i, x_i)_t + \alpha \cdot \left(\beta \cdot \vec{v}^{\ell^*}_{\text{T}} + (1-\beta) \cdot \vec{v}^{\ell^*}_{\text{C}}\right) \tag{6}$$

where α controls the overall intervention strength, and β determines the relative weight of each steering vector. This intervention guides the model's representations toward more faithful, image-grounded reasoning. As shown in Section 5.2, our approach effectively reduces both types of hallucinations while preserving overall response quality.

5 Experiment

5.1 Experiment Setup

Benchmarks and Evaluation Metrics (1) **POPE** (Li et al., 2023) is a dedicated benchmark for evaluating object hallucination. It probes a model's ability to recognize specific objects by posing binary questions. Performance is measured using standard classification metrics, including Accuracy, Precision, Recall, and F1 score. (2) **MME** (Fu et al., 2024) construct paired questions with opposite answers ("yes" and "no") for each image. In addition to question-level accuracy (Acc), it introduces image-level accuracy (Acc+), which requires both answers to be correct. The final performance is measured by the MME Score, calculated as the sum of Acc and Acc+. (3) **CHAIR** (Rohrbach et al., 2018) assesses hallucinations in image captioning by analyzing model-generated captions. It measures the proportion of object mentions that do not appear in the ground-truth annotations. The evaluation includes three metrics: CHAIR_i for instance-level and CHAIR_s for sentence-level hallucination, and Recall for (4) **AMBER** (Wang et al., 2023) is evaluates object, attribute, and relation hallucinations in both discriminative and generative tasks. For discriminative tasks, it adopts standard metrics such as Accuracy, Precision, Recall, and F1 score. In generative settings, it employs CHAIR_i, Cover, Hal, and Cog. A detailed explanation of these metrics can be found in Section B.2. To enable holistic assessment, AMBER also introduces a unified metric, defined as:

$$\text{AMBER Score} = \frac{1}{2} \times (1 - \text{CHAIR}_i + \text{F1}). \tag{7}$$

Setting	Method	LLaVA-v1.5				QwenVL				LLaVA-Next			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
random	Sampling	83.8	82.4	86.1	84.2	84.9	96.0	72.9	82.9	84.4	94.7	72.8	82.3
	VCD	**85.0**	82.7	86.1	84.2	85.5	96.0	71.1	83.6	86.0	96.5	74.8	84.3
	VTI	83.0	80.6	86.8	83.6	85.3	95.1	73.8	83.5	84.8	94.0	74.4	83.1
	SHARP	**85.0**	83.8	86.9	**85.3**	**86.1**	97.4	74.1	**84.2**	**88.4**	93.5	82.6	**87.7**
popular	Sampling	82.0	79.7	85.9	82.6	84.0	94.7	72.1	81.9	83.2	90.9	73.8	81.5
	VCD	82.1	78.5	88.3	**83.2**	84.9	94.5	74.9	83.6	84.5	92.9	74.8	82.9
	VTI	80.4	76.4	88.1	81.8	83.0	93.3	74.1	82.3	81.8	82.5	80.8	81.6
	SHARP	**82.3**	79.4	87.3	**83.2**	**85.8**	96.3	74.5	**84.0**	**85.9**	88.9	82.0	**85.3**
adversarial	Sampling	75.8	71.3	86.3	78.1	82.1	90.0	72.3	80.2	79.5	84.1	72.9	78.1
	VCD	76.3	71.5	87.3	78.7	**84.0**	90.6	74.9	82.0	80.9	85.2	74.8	79.7
	VTI	76.0	70.7	88.8	78.8	83.2	91.1	74.5	81.8	79.0	81.6	74.9	78.1
	SHARP	**76.8**	72.3	86.9	**79.0**	83.9	92.6	73.7	**82.1**	**82.8**	82.7	82.9	**82.8**

Table 1: **Results on POPE-MSCOCO benchmark.** "Acc", "Pre", "Rec", and "F1" stand for Accuracy, Precision, Recall, and F1 score, respectively. The reported results are derived from the MS-COCO dataset. The best **Acc** and **F1** scores for each setting and model are bolded.

Model	Method	Object-level		Attribute-level		Total
		Exist	Count	Pos	Color	
LLaVA-v1.5	Sampling	170.0	103.3	108.3	128.3	510.0
	VCD	180.0	110.0	108.3	133.3	531.7
	VTI	190.0	138.3	131.7	145.0	605.0
	SHARP	175.0	155.0	103.3	180.0	**613.3**
Qwen-VL	Sampling	160.0	143.3	113.3	165.0	581.7
	VCD	165.0	140.0	113.3	175.0	593.3
	VTI	182.0	125.0	118.0	143.0	568.0
	SHARP	175.0	145.0	103.3	185.0	**608.3**
LLaVA-Next	Sampling	175.0	143.3	131.7	145.0	595.0
	VCD	190.0	145.0	116.7	160.0	611.7
	VTI	180.0	110.0	101.7	140.0	531.7
	SHARP	195.0	128.3	143.3	165.0	**631.7**

Table 2: **Results on MME benchmark.** The performance is measured by MME Score. The "Total" column represents the sum of four individual results in each row. The best **Total** score for each model is bolded.

Baselines We compare SHARP with various hallucination mitigation methods for LVLMs, as outlined below: VCD (Leng et al., 2024) suppresses language priors by subtracting prior-induced noise to enhance visual grounding during decoding. VTI (Liu et al., 2024e) reduces hallucinations by steering latent space representations during inference to enhance the stability of vision features.

Implementation details We evaluate our proposed approach on three recent large vision-language models (LVLMs): LLaVA-1.5 (Liu et al., 2023), Qwen-VL (Bai et al., 2023), and LLaVA-NEXT (Liu et al., 2024d). For all three models, we set the representation steering layer ℓ^* to the 10th layer and fix the intervention strength at $\alpha = 5$. We adopt a sampling-based decoding strategy for both SHARP and all state-of-the-art baseline methods. For baselines, we follow the configurations reported in their original papers and released code to ensure fair comparison. Additional details are provided in Appendix B.3.

5.2 Main Results

Results on POPE The overall performance of our proposed method SHARP on POPE benchmark on the MSCOCO dataset is shown in Table 1. From the table, we can observe that SHARP consistently outperforms all baselines across the three evaluation settings. It significantly improves upon the base sampling strategy, as the activation steering guides the model's internal representations toward factual concepts, leading to more reliable responses. Moreover, SHARP achieves the highest overall performance and consistently surpasses strong baselines. In contrast, methods such as VCD and VTI lack the robustness and stability demonstrated by our approach. Complete results on the A-OKVQA and GQA datasets are provided in Appendix C.1.

Results on MME To accurately evaluate whether a model truly understands an image, MME adopts a balanced evaluation protocol: an image is considered correctly understood only if the model answers both the corresponding "yes" and "no" questions correctly. Following prior work (Leng et al., 2024),

Model	Method	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Recall	Length
LLaVA-v1.5	Sampling	52.8	15.9	**77.3**	93.4
	VCD	51.0	14.9	77.2	101.9
	VTI	36.8	13.6	66.6	66.7
	SHARP	**34.8**	**10.6**	59.7	93.4
Qwen-VL	Sampling	2.8	3.0	31.0	5.3
	VCD	1.4	1.2	30.8	4.0
	VTI	1.9	1.9	30.2	4.5
	SHARP	**0.8**	**0.8**	**32.4**	3.7
LLaVA-Next	Sampling	35.8	12.0	59.5	179.0
	VCD	40.2	10.7	**62.1**	171.2
	VTI	**26.4**	**7.7**	58.5	187.6
	SHARP	31.8	9.2	61.5	180.5

Table 3: **Results on CHAIR benchmark.** Lower $CHAIR_i$ and $CHAIR_s$, along with higher Recall, correspond to better performance. The best **$CHAIR_s$**, **$CHAIR_i$**, and **Recall** scores for each setting and model are bolded, and the second-best scores are underlined.

we evaluate our method on the object-level and attribute-level perception tasks in MME, as shown in Table 2. Our approach outperforms state-of-the-art methods, achieving improvements of 8.3, 15, and 20 points across the three LVLMs respectively. However, Qwen-VL shows a slight drop on the position subset and LLaVA-NEXT on the count subset, likely due to the models' limited spatial and numerical reasoning—capabilities less impacted by activation steering. We leave further investigation to future work.

Results on CHAIR Beyond the relatively simplified and evaluation-friendly discriminative tasks, generative tasks are of greater practical importance, as LVLMs are predominantly used for content generation. As shown in Table 3, SHARP significantly reduces both $CHAIR_s$ and $CHAIR_i$, indicating that it not only decreases the number of hallucinated objects in generated responses but also improves the proportion of factually grounded captions. It is worth noting that Qwen-VL inherently exhibits fewer hallucinations due to its tendency to produce shorter responses. Nevertheless, our method is still able to further mitigate hallucinations in this model, demonstrating its generalizability. Additionally, we report *Recall* and *Length* as supplementary metrics to assess the richness and informativeness of generated content. Prior methods often reduce hallucinations at the cost of lower recall and shorter outputs, leading to overly conservative generations. Unlike prior methods that suppress hallucinations at the cost of expressiveness, our approach improves both, indicating effective and precise mitigation without sacrificing generation quality.

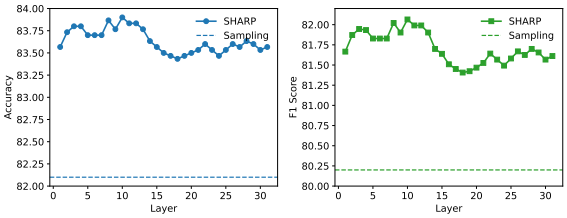


Figure 3: The ablation results of target layer ℓ^* for Qwen-VL model on the POPE adversarial subset.

Results on AMBER The AMBER benchmark offers a comprehensive evaluation across both discriminative and generative tasks. The results are summarized in Table 4. For the discriminative tasks, SHARP achieves the highest performance across all evaluation metrics, with an average absolute improvement of 5.07% in accuracy and 4.53% in F1 score. In the generative setting, SHARP consistently improves all metrics across all evaluated models. Specifically, it effectively reduces hallucination-related metrics such as $CHAIR_s$ and Hal., while simultaneously increasing Cover, reflecting a broader and richer content generation. Although VTI achieves slightly lower hallucination scores on LLaVA-1.5, it does so at the expense of significantly reducing Cover, which compromises the informativeness and diversity of generated outputs. Finally, SHARP achieves the highest overall AMBER Score across all models, outperforming the strongest baselines by a margin of 1.20%. These results demonstrate the generalizability and effectiveness of our approach—it not only mitigates hallucinations but also preserves the generative capacity and utility of LVLMs.

5.3 Analysis

Ablation on target layer ℓ^* In this section, we conduct experiments on POPE under adversarial setting to analyze the performance fluctuation of SHARP by the steering layer ℓ^*. The steering layer ℓ^* refers to the transformer layer where the steering vectors are injected, ranging from 0 to 31 across all three models. As shown in Fig. 3, SHARP consistently improves performance across different layers, demonstrating strong robustness to the choice of ℓ^*. Notably, the best results are achieved in the middle layers (10–15). This aligns with our probing analysis in Sec. 3, which shows that representation separability increases up to these layers before stabilizing—suggesting that mid-level layers begin to encode cause-distinguishable features

Model	Method	Discriminative				Generative				AMBER Score ↑
		Acc	Pre	Rec	F1	CHAIR$_i$↓	Cover↑	Hal↓	Cog*	
LLaVA-v1.5	Sampling	67.0	85.2	60.9	71.0	12.0	50.3	51.0	4.6	79.5
	VCD	67.3	86.1	60.5	71.1	10.1	51.2	43.6	4.3	80.6
	VTI	66.5	84.5	60.6	70.6	**6.9**	47.2	**27.0**	1.8	81.9
	SHARP	**74.2**	89.9	68.8	**77.9**	8.5	**52.1**	39.2	4.8	**84.7**
Qwen-VL	Sampling	82.9	88.0	85.9	86.9	4.8	31.3	9.2	0.3	91.1
	VCD	84.1	89.2	86.6	87.9	3.5	35.2	7.7	0.3	92.2
	VTI	83.5	88.3	86.1	87.3	3.1	33.8	6.8	0.25	92.4
	SHARP	**84.4**	89.1	87.1	**88.1**	**2.7**	**36.1**	**5.7**	0.2	**92.7**
LLaVA-Next	Sampling	72.9	82.4	75.2	78.6	12.0	56.5	59.6	5.1	83.3
	VCD	74.3	83.9	75.8	79.6	11.8	59.1	58.6	5.0	83.9
	VTI	75.2	79.3	84.7	81.9	10.3	58.5	58.3	4.9	87.1
	SHARP	**79.4**	86.5	81.8	**84.1**	**8.9**	**61.5**	**50.5**	4.8	**87.6**

Table 4: **Results on AMBER benchmark.** In discriminative tasks, "Acc", "Pre", "Rec", and "F1" stand for Accuracy, Precision, Recall, and F1 score, respectively. Higher values for these metrics indicate superior performance. In generative tasks, lower CHAIR$_i$ and Hal, along with higher Cover, signify better performance. *Cog measures the extent to which hallucinations in LVLMs align with human cognition and is therefore not directly comparable. The best **Acc**, **F1**, **CHAIR$_i$**, **Hal**, and **AMBER Score** for each model are bolded

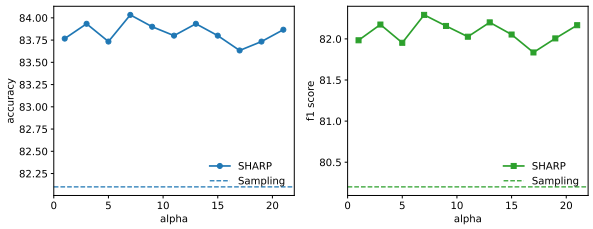


Figure 4: The ablation results of intervention strength α for Qwen-VL on the POPE adversarial subset.

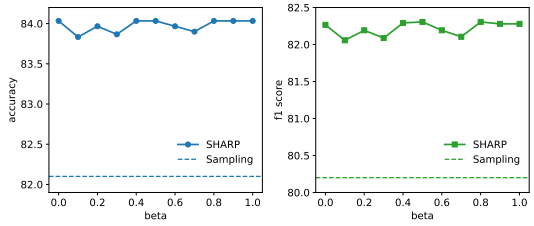


Figure 5: The ablation results of steering vector weight β for the two causes — textual prior and visual-context conflict — evaluation results for Qwen-VL on the POPE adversarial subset.

that influence downstream representations. Based on this observation, we select the 10th layer as the default intervention layer, which achieves stable and generalized improvements across tasks.

Ablation on intervention strength α The intervention strength α controls the magnitude of steering applied to the internal representations. We vary the intervention strength α from 1 to 21, as shown in Fig. 4. The results demonstrate that SHARP consistently remains effective across this spectrum, outperforming the backbone model under all settings. This indicates that SHARP is not only powerful in mitigating hallucinations but also robust to variations in the hyperparameter α, making it practical and adaptable for real-world use without needing fine-tuned adjustments.

Ablation on steering vector weight β We explore the effect of the relative weight β assigned to the two cause-specific steering vectors $\vec{v}_{\mathrm{T}}^{\ell^*}$ and $\vec{v}_{\mathrm{C}}^{\ell^*}$, as shown in Fig. 5. Adjusting β controls the emphasis between different hallucination causes, enabling us to understand the contribution of each direction and optimize the intervention for comprehensive hallucination mitigation. When β is set to 0 or 1—i.e., only one vector is applied—the performance already surpasses the baseline, demonstrating the effectiveness of each cause-specific intervention. As β varies between 0 and 1, the model shows some fluctuations but maintains a consistently strong performance, with the best results observed around $\beta = 0.4$. These findings highlight the robustness of our approach and the benefit of combining both steering directions for comprehensive hallucination mitigation.

6 Conclusion

In this work, we present a fine-grained analysis of hallucination in LVLMs by disentangling two key causes: over-reliance on textual priors and visual-context conflict. Our findings reveal that LVLMs internally encode signals indicative of these causes, enabling targeted representation-level intervention. Building on this insight, we propose SHARP, a novel inference-time approach that steers internal activations using cause-specific vectors without requiring model retraining. Comprehensive experiments on three LVLMs over four benchmarks show our SHARP can significantly reduces hallucinations across tasks while preserving generation quality.

Limitations

While SHARP demonstrates significant advantages in effectiveness, several limitations remain that warrant further investigation:

First, although our disentangled hallucination mechanisms and steering vector derivation are efficient and broadly applicable, they may not achieve optimal disentanglement across all hallucination scenarios. The current contrastive approach relies on the model's inherent capacity to separate hallucination-inducing features via simple activation arithmetic. Future work could incorporate causal analysis to uncover a more comprehensive set of hallucination triggers, enabling finer-grained categorization and allowing for the merging or decomposition of steering vectors based on distinct causes.

Second, our current implementation adopts a relatively simple strategy for deriving steering vectors. Exploring more advanced or targeted strategies could help identify more precise intervention directions, potentially improving the overall effectiveness of hallucination mitigation. We leave these explorations for future work.

Lastly, while our method currently focuses on mitigating hallucinations in VQA and image captioning, its impact on long-form or complex reasoning appears limited, as shown in Sec. C.4. This is expected, since SHARP primarily targets visually grounded hallucinations rather than general reasoning capabilities. Future work will investigate how to extend our approach to more complex and diverse hallucination scenarios, aiming to enhance overall performance.

Ethical Considerations

While our approach aims to reduce hallucinations and improve the reliability of LVLMs, it does not explicitly address potential biases in training data or the risk of downstream misuse of steering interventions. We underscore the importance of responsible deployment and fairness to ensure that enhanced controllability is leveraged to improve model safety rather than manipulate content or reinforce harmful biases. Our experiments are conducted using publicly available pre-trained models, including LLaVA-v1.5, Qwen-VL, LLaVA-NEXT, and the GPT-4o-mini API. All models and datasets have been carefully curated by their original authors to mitigate potential ethical concerns.

Acknowledgements

This work is supported by National Natural Science Foundation of China (62576339, 62372454).

References

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. *Preprint*, arXiv:2308.12966.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2616–2627.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024a. Unified hallucination detection for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3235–3252.

Xinlong Chen, Yuanxing Zhang, Qiang Liu, Junfei Wu, Fuzheng Zhang, and Tieniu Tan. 2025. Mixture of decoding: An attention-inspired adaptive decoding strategy to mitigate hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8525–8542, Vienna, Austria. Association for Computational Linguistics.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024b. Halc: Object hallucination reduction via adaptive focal-contrast decoding. In *Forty-first International Conference on Machine Learning*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models.

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*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

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.

Junho Kim, Yeonju Kim, and Yong Man Ro. 2024. What if...?: Counterfactual inception to mitigate hallucination effects in large multimodal models. *CoRR*.

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.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. *ICLR*.

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

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024d. Llava-next: Improved reasoning, ocr, and world knowledge.

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

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024e. Reducing hallucinations in vision-language models via latent space steering. *arXiv preprint arXiv:2410.15778*.

Xiaoyuan Liu, Wenxuan Wang, Youliang Yuan, Jentse Huang, Qiuzhi Liu, Pinjia He, and Zhaopeng Tu. 2024f. Insight over sight? exploring the vision-knowledge conflicts in multimodal llms. *arXiv preprint arXiv:2410.08145*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

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*.

A Paszke. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

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, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *Annual Meeting of the Association for Computational Linguistics*.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. 2025. MLLM can see? dynamic correction decoding for hallucination mitigation. In *The Thirteenth International Conference on Learning Representations*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and 1 others. 2023. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Zhecan Wang, Garrett Bingham, Adams Wei Yu, Quoc V Le, Thang Luong, and Golnaz Ghiasi. 2024. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. In *European Conference on Computer Vision*, pages 288–304. Springer.

Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. 2024. Logical closed loop: Uncovering object hallucinations in large vision-language models. *arXiv preprint arXiv:2402.11622*.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. 2024. Mitigating hallucination in large vision-language models via modular attribution and intervention. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2025. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. *ICLR*.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, and 1 others. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

A Algorithm

The complete procedure of our SHARP is formalized in Algorithm 1.

Algorithm 1: SHARP: Steering Hallucinations via Representation Processing

Input: Base LVLM M; Target layer ℓ^*; Scaling factor α; Blending weight β; VQA dataset (x, q)

Output: Steered faithful LVLM $\mathbf{M}$

Stage 1: Cause-Specific Dataset Construction;

foreach (x_i, q_i) *in benchmark dataset* **do**

  Generate answer $\hat{y}_i \leftarrow M(x_i, q_i)$;

  Score factuality $s_i \leftarrow \text{Judge}(y_i^{\text{ref}}, \hat{y}_i)$;

  Classify $(x_i, q_i, \hat{y}_i)$ into $\mathcal{D}^{(m)}_{\text{faithful}}$ or $\mathcal{D}^{(m)}_{\text{hallucinated}}$ based on s_i and cause $m \in \{\text{T}, \text{C}\}$;

Stage 2: Cause-Specific Steering Vector Extraction;

foreach *cause* $m \in \{T, C\}$ **do**

  Compute cause vector at layer ℓ^*:

$$\vec{v}_m^{\ell^*} = \frac{\sum_{(v_i,x_i)\in\mathcal{D}^{(m)}_{\text{faithful}}} h^{\ell^*}(v_i, x_i)}{|\mathcal{D}^{(m)}_{\text{faithful}}|} - \frac{\sum_{(v_j,x_j)\in\mathcal{D}^{(m)}_{\text{hallucinated}}} h^{\ell^*}(v_j, x_j)}{|\mathcal{D}^{(m)}_{\text{hallucinated}}|} \tag{8}$$

Stage 3: Dynamic Activation Steering During Inference;

Compute adaptive steering vector: $\vec{v}^{\ell^*}_{\text{adaptive}} = \beta \cdot \vec{v}^{\ell^*}_{\text{T}} + (1-\beta) \cdot \vec{v}^{\ell^*}_{\text{C}}$;

foreach *token position* t *during generation* $(t \geq |x_i|)$ **do**

  Retrieve hidden state $h_t^{\ell^*} \leftarrow M(x, q)_t$;

  Inject: $h^{\ell^*}_{\text{steered},t} = h_t^{\ell^*} + \alpha \cdot \vec{v}^{\ell^*}_{\text{adaptive}}$;

B Experiment Details

B.1 Cause-related Stimulation Data Construction Details

HaloQuest Dataset HaloQuest dataset (Wang et al., 2024) is a VQA dataset designed to various aspects of multimodal hallucination such as false premises, insufficient contexts, and visual challenges. Here we use the false premises split and insufficient contexts split to construct our cause-related stimulation data. Specifically, samples in the false premise data split contain statements or assumptions that directly contradict the visual content of the image. They are designed to test whether the model can correctly prioritize visual evidence over misleading linguistic cues. Besides, samples in the insufficient context cannot be definitively answered based on the image alone. They probe whether models will resort to biases or unfounded assumptions instead of acknowledging the limits of the provided information. The statistical information of the dataset is shown in Table 5

The prompt used for scoring The prompt template we use to score LVLM's output on the cause-related stimulation dataset is shown in Fig. 6

B.2 Metrics

Metrics on CHAIR. CHAIR evaluates hallucinations in image captioning by analyzing the objects mentioned in model-generated captions. It quantifies the proportion of object mentions that are not present in the ground-truth annotations. Two variants are considered: CHAIR_i for instance-level and CHAIR_s for sentence-level hallucination. An additional metric, Recall, measures the fraction of ground-truth objects that are correctly mentioned.

$$\text{CHAIR}_i = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|} \tag{9}$$

$$\text{CHAIR}_s = \frac{|\{\text{sentences with hallucinations}\}|}{\{\text{all sentences}\}} \tag{10}$$

$$\text{Recall} = \frac{|\{\text{accurately mentioned objects}\}|}{|\{\text{ground-truth objects}\}|} \tag{11}$$

Metrics on AMBER. In the context of generative tasks, AMBER employs CHAIR_i (as defined in Equation 9), Cover, Hal, and Cog as comprehensive metrics. For clarity, we represent the objects mentioned in a model's response as $R_{\text{obj}} = \{r_1, r_2, \ldots, r_m\}$, and the ground-truth objects annotated in the image as $G_{\text{obj}} = \{g_1, g_2, \ldots, g_n\}$.

Cover: Measures the proportion of ground-truth objects correctly mentioned.

$$\text{Cover} = \frac{|R_{\text{obj}} \cap G_{\text{obj}}|}{|G_{\text{obj}}|} \tag{12}$$

Dataset Splits	Train	Eval	Total
False Premise Questions	2,698	304	3,002
Questions with Insufficient Context	1,469	121	1,590

Table 5: Information about the two data splits in HaloQuest dataset.

System prompt

{default system prompt of GPT-4o-mini}

Prompt

You are a helpful assistant.
You will be given a pair of answers:
[Ground Truth Answer]: the human-annotated correct answer
[Model Answer]: the answer generated by a model
Your task is to judge whether the model answer is correct based on the ground truth answer.

Evaluation criteria:
If the model answer accurately conveys the core information of the ground truth answer, even if phrased differently, it should be judged as Correct.
If the model answer contains factual errors, omits key information, or is inconsistent with the meaning of the ground truth answer, it should be judged as Incorrect.
Focus solely on the factual correctness of the content. Do not evaluate language fluency or style.
Output format:
correct / incorrect

[Ground Truth Answer]
{gt_answer}
[Model Answer]
{model_output}

Figure 6: The prompt template we use to score LVLM's answers on open ended VQA

Hal: Indicates whether any hallucinated object appears in the response.

$$\text{Hal} = \begin{cases} 1, & \text{if CHAIR} \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

Cog: Measures the extent to which hallucinated objects align with those frequently hallucinated by humans. Let $H_{\text{obj}} = \{h_1, h_2, \ldots, h_p\}$ denote the set of objects humans are prone to hallucinate. Cog is calculated as:

$$\text{Cog} = \frac{|R_{\text{obj}} \cap H_{\text{obj}}|}{|R_{\text{obj}}|} \tag{14}$$

B.3 Implementation Details

In this work, all experiments were conducted using the PyTorch framework (Paszke, 2019). All baseline LVLMs and hallucination mitigation methods were re-implemented following their original publications. We retained the default hyperparameter settings for all backbone LVLMs and baseline methods. Experiments were performed on 8 NVIDIA RTX 3090 GPUs (24 GB each), and all reported results are based on a single run. In our experiments, we employ the following library versions: Transformers 4.40.0 and scikit-learn 1.2.2.

C Full Evaluation Results

C.1 Additional Results on POPE Benchmark

In addition to the MSCOCO results in Table 1, we report POPE performance on the AOKVQA and GQA subsets (Tables 6 and 7). Each subset comprises 500 images, each paired with 6 questions. Negative samples are generated under three settings: random (randomly selected absent objects), popular (frequently occurring but absent objects), and adversarial (contextually co-occurring yet absent objects). The numbers of positive and negative samples are kept balanced. It can be seen that SHARP consistently achieves strong performance across both subsets and all sampling settings.

C.2 Ablation Studies

Ablation on target layer ℓ^* The ablation study results of ℓ^* on Llava-next model and LLaVA-V1.5 model are shown in Fig. 7 and Fig. 8

Ablation on intervention strength α The ablation study results of α on LLaVA-V1.5 model and Llava-next model are shown in Fig. 9 and Fig. 10.

Ablation on steering vector weight β The ablation study results of β for the other two models are show in Fig. 11 and Fig. 12.

Setting	Method	LLaVA-V1.5				Qwen-VL				LLaVA-Next			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
random	Sampling	81.8	76.4	92.1	83.5	86.8	93.2	79.5	85.8	83.8	87.2	79.2	83.0
	VCD	81.2	75.2	93.0	83.2	87.4	92.9	81.1	86.6	84.8	89.2	79.3	83.9
	VTI	83.0	80.6	86.8	83.6	85.0	89.0	80.0	85.3	84.8	87.3	81.3	84.2
	SHARP	**83.2**	77.7	93.1	**84.7**	**87.8**	94.0	80.8	**86.9**	**89.0**	88.3	89.8	**89.1**
popular	Sampling	75.3	69.1	91.5	78.7	85.6	90.6	79.5	84.7	81.4	83.4	78.3	80.8
	VCD	74.7	68.2	92.5	78.5	86.3	89.5	81.2	85.1	81.5	82.6	79.9	81.2
	VTI	76.1	69.5	92.9	79.5	82.5	87.0	79.8	83.3	80.4	82.7	76.9	79.7
	SHARP	**78.2**	71.6	93.4	**81.1**	**87.0**	92.1	80.9	**86.1**	**84.7**	81.7	89.3	**85.3**
adversarial	Sampling	67.4	61.8	91.2	73.7	80.4	80.1	80.9	80.5	73.2	71.0	78.4	74.5
	VCD	68.1	61.9	93.8	74.6	80.7	80.1	81.6	80.8	74.7	72.2	80.3	76.0
	VTI	68.2	62.1	93.6	74.6	74.5	78.0	74.1	75.0	72.8	70.4	78.6	74.3
	SHARP	**68.8**	62.7	93.3	**75.0**	**81.3**	81.5	81.0	**81.2**	**76.9**	71.5	89.5	**79.5**

Table 6: **Results on POPE-AOKVQA benchmark.** "Acc", "Pre", "Rec", and "F1" denote Accuracy, Precision, Recall, and F1 score, respectively. The best **Acc** and **F1** scores for each model under each prompting setting are bolded.

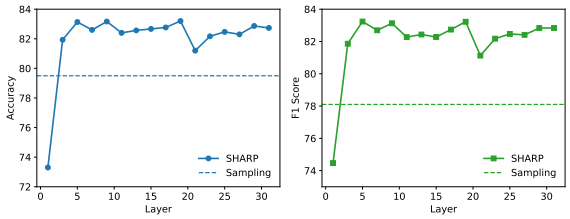


Figure 7: The layer ℓ^* ablation results for Llava-Next model on POPE-MSCOCO benchmark under the adversarial evaluation setting

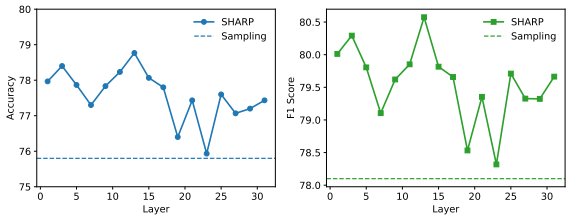


Figure 8: The layer ℓ^* ablation results for LLaVA-V1.5 model on POPE-MSCOCO benchmark under the adversarial evaluation setting

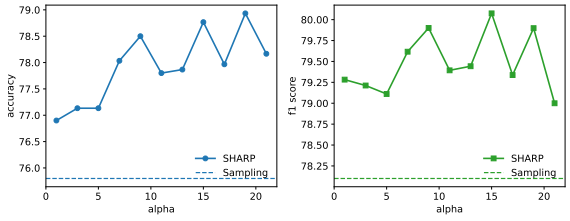


Figure 9: The intervention strength α ablation results for LLaVA-V1.5 model on POPE-MSCOCO benchmark under the adversarial evaluation setting

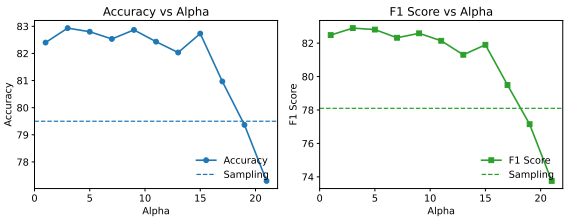


Figure 10: The intervention strength α ablation results for Llava-Next model on POPE-MSCOCO benchmark under the adversarial evaluation setting

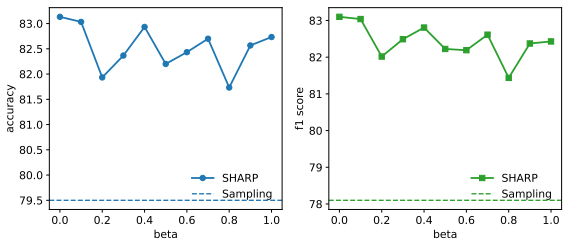


Figure 11: The steering vector weight β ablation results for Llava-Next model on POPE-MSCOCO benchmark under the adversarial evaluation setting

Setting	Method	LLaVA-v1.5-7B				Qwen-VL				LLaVA-Next			
		Acc	Pre	Rec	F1	Acc	Pre	Rec	F1	Acc	Pre	Rec	F1
random	Sampling	81.6	75.6	93.2	83.5	81.3	88.8	71.5	79.2	83.1	85.8	79.4	82.5
	VCD	82.2	76.0	94.1	84.1	82.0	87.6	74.5	80.5	83.4	87.0	80.1	83.4
	VTI	79.8	73.3	93.9	82.3	82.0	85.0	73.9	82.4	83.3	85.9	79.6	82.6
	SHARP	**83.2**	77.8	92.9	**84.7**	**83.7**	90.7	75.2	**82.2**	**87.3**	86.8	88.0	**87.4**
popular	Sampling	73.1	66.7	92.5	77.5	75.9	78.1	72.0	74.9	78.5	78.7	78.2	78.5
	VCD	71.5	64.7	94.5	76.8	75.9	76.6	74.7	75.6	78.2	77.2	80.1	78.6
	VTI	73.5	66.7	94.1	78.0	76.5	77.0	73.5	77.9	78.6	79.9	76.5	78.2
	SHARP	**73.7**	66.8	93.9	**78.1**	**79.9**	83.2	75.0	**78.9**	**80.8**	76.9	88.0	**82.1**
adversarial	Sampling	68.0	62.0	93.4	74.5	75.5	77.8	71.2	74.4	73.3	71.3	78.0	74.5
	VCD	67.6	61.5	94.4	74.5	76.7	77.8	74.7	76.2	74.2	71.8	79.8	75.6
	VTI	68.0	61.9	93.9	74.6	71.0	72.5	73.2	74.5	73.2	71.4	77.5	74.3
	SHARP	**69.4**	63.1	93.7	**75.4**	**79.2**	82.0	74.9	**78.3**	**76.6**	71.8	87.4	**78.9**

Table 7: **Results on POPE-GQA benchmark.** "Acc", "Pre", "Rec", and "F1" denote Accuracy, Precision, Recall, and F1 score, respectively. The best **Acc** and **F1** scores for each model under each prompting setting are bolded.

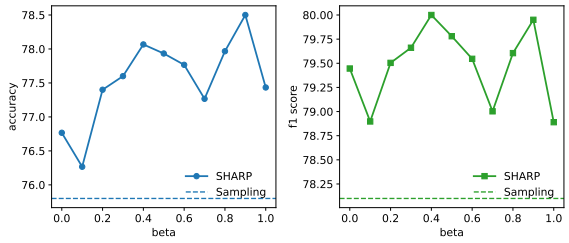


Figure 12: The steering vector weight β ablation results for LLaVA-V1.5 model on POPE-MSCOCO benchmark under the adversarial evaluation setting

Model	Method	ACC	F1
LLaVA-Next	Sampling	79.5	78.1
	VCD	80.9	79.7
	VTI	79.0	78.1
	SHARP (HaloQuest)	**82.8**	82.8
	SHARP (ConflictVis)	**82.8**	**82.9**

Table 8: Results of generalization experiments by replacing the textual-prior subset with ConflictVis (Liu et al., 2024f) on the POPE adversarial subset.

C.3 Generalization Experiments with an Alternative Textual-Prior Dataset

To evaluate the generalizability of our approach, we replaced the textual-prior subset from HaloQuest with ConflictVis (Liu et al., 2024f), a dataset specifically designed to induce hallucinations through counterfactual visual content and counter-commonsense queries. ConflictVis comprises 374 original images and 1,122 high-quality QA pairs, which often trigger hallucinations or unwarranted denials caused by textual bias.

As shown in Table 8, SHARP maintains superior performance even when the steering vectors are derived from ConflictVis instead of HaloQuest, underscoring the robustness and transferability of our method across datasets.

C.4 Generalization on Reasoning Benchmarks

To assess whether SHARP affects general reasoning, we evaluate it on on two challenging reasoning benchmarks MMMU (Yue et al., 2024) and MathVista (Lu et al., 2024) (Table 9). Results show only minor fluctuations compared to standard sampling, indicating that SHARP preserves reasoning abilities without enhancing them. This is expected, as SHARP primarily targets visually grounded hallucinations from textual priors or misleading prompts, which are largely independent of broader reasoning. Future work will explore additional hallucination causes and corresponding intervention vectors to jointly mitigate hallucinations and enhance LVLM capabilities.

C.5 Inference-Time Efficiency Analysis

While inference-time strategies reduce hallucinations, they inevitably introduce computational overhead. We measure the average time per response on the CHAIR benchmark (Table 10). Sampling-based decoding is fastest and serves as the baseline, while VCD incurs substantial overhead due to dual forward passes. In contrast, VTI and our SHARP achieve a better balance of performance and effi-

Model	Method	MMMU ↑	MathVista ↑
LLaVA-V1.5	Sampling	33.8	22.6
	SHARP	33.3	22.8
Qwen-VL	Sampling	31.7	34.7
	SHARP	31.9	34.1
LLaVA-Next	Sampling	34.8	31.3
	SHARP	34.7	31.9

Table 9: Performance on general reasoning benchmarks.

Model	Method	Seconds / Sample
LLaVA-V1.5	Sampling	3.68
	VCD	8.93
	VTI	4.24
	SHARP	4.18
LLaVA-Next	Sampling	6.80
	VCD	13.97
	VTI	7.66
	SHARP	7.04

Table 10: Average response generation time on the CHAIR benchmark.

ciency, with SHARP adding only minimal latency by intervening in few positions. This shows that SHARP is both effective and practical for real-world use.

D Case Studies

There are two more case results, shown in Fig. 13 and Fig. 14.

E Licensing

The POPE, MME, CHAIR, and AMBER datasets are released for academic use and are specifically designed for hallucination detection tasks. Our usage of these datasets aligns with their intended purpose. LLaVA-v1.5-7B and LLaVA-NEXT-7B are released under the Apache 2.0 License, a permissive open-source license. Qwen-VL-7B is released under the Tongyi Qianwen LICENSE AGREEMENT by Alibaba Cloud, which permits non-commercial academic usage and redistribution under certain conditions. Our use of these models is limited to academic research and complies with their respective license terms.

Text input: Please describe this image in detail.

Sampling: In the image, a young baseball player is captured in the midst of a powerful swing. The player, dressed in a black t-shirt and gray pants, grips a vibrant red and blue baseball bat in their hands. They are positioned in the batter's box, a black area marked on a dirt field, preparing to hit an incoming ball. \n\nThe player's helmet, a safety measure against fast-pitched balls, is a stark contrast against their dark attire. The helmet and the player's intense focus suggest they are fully invested in the game.The field is enclosed by a brick wall on one side and a fence on another, providing a boundary for the player's swing. The image does not just capture a moment of action, but also the anticipation and concentration required in the sport of baseball.

SHARP: In the heart of the image, a young baseball player, dressed in a black t-shirt and gray pants, stands poised on a dirt field. His stance is one of anticipation, his body coiled in readiness for the pitch. His right hand grips a wooden bat, its grain and texture a testament to countless games played. The player's helmet, a stark contrast to his black t-shirt, sits atop his head, a silent sentinel protecting his head from the high-velocity pitches that are about to be thrown. His eyes are focused on the pitcher, his gaze unwavering as he prepares to swing.\n\nThe field itself is a study in contrasts - the dirt ground beneath his feet, the grassy area that lies beyond, and the brick building in the background. The building's red brick facade adds a touch of color to the scene, its rustic charm complementing the modernity of the baseball field.

Figure 13: Case study of generative tasks using LLaVA-1.5. We compare responses generated by sampling and our proposed SHARP. Hallucinated content is highlighted in red, while more detailed and accurate content is marked in green.

Text input: Please describe this image in detail.

Sampling: Sure! The image captures a small, humble bathroom in the midst of a renovation process. The walls are a pristine white, contrasting with the dark floor made of shiny tiles. A toilet, white in color like the walls, is mounted on the left side of the room against a wall. A white pipe runs along the floor, presumably for the toilet flush system. On the right side of the room, a white trash can is placed on the floor. A white staircase with wooden steps can be seen in the background, suggesting that the bathroom is located on an upper floor in a residential home. The room gives off an impression of being newly furnished and ready for occupancy.

SHARP: The image captures a scene of a bathroom in the midst of a renovation. The room is dominated by a pristine white toilet, which is the only object in the room. The toilet is positioned in the center of the room, at the right side of the image. The walls of the bathroom are painted in a stark white, providing a clean and minimalist aesthetic. The floor, however, is a different story. It's covered in white tiles, but they are not uniformly laid out. Some tiles are missing, revealing the underlying floor beneath. On the right side of the image, there's a wooden staircase. Its presence adds a touch of warmth to the otherwise cold and sterile environment. The staircase is partially visible, with only the first few steps visible in the frame. The image is taken from a low angle, looking up at the toilet. This perspective emphasizes the height of the toilet and gives a sense of the room's dimensions. Despite the ongoing renovation, the bathroom maintains a sense of order and cleanliness.

Figure 14: Case study of generative tasks using LLaVA-Next. We compare responses generated by sampling and our proposed SHARP. Hallucinated content is highlighted in red, while more detailed and accurate content is marked in green.