RIVS: Mitigating Hallucination in Large Vision-Language Models via Representation Intervention on Visual Grounding Shift

Xuanyu Yin[1,2], **Xiaoye Qu**[1,2], **Wei Wei**[*1,2]
[1]Cognitive Computing and Intelligent Information Processing (CCIIP) Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology
[2]Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL)
{xuanyu, xiaoye, weiw}@hust.edu.cn

Abstract

Large Vision-Language Models (LVLMs) demonstrate powerful generative capabilities yet remain prone to object hallucinations. Most existing methods mitigate this issue through training or decoding strategies, but provide limited exploration of how hallucinations arise from internal representations during generation. In this work, we study hallucination from the perspective of dynamic representation shift during generation and propose **R**epresentation **I**ntervention based on **V**isual Grounding **S**hift (**RIVS**). Specifically, we first design an adaptive threshold-based method to identify visual attention drop points within the generation, finding that hallucinated tokens usually occur with decay of visual attention. Based on this observation, we construct a hallucination-related subspace from representation differences around these points on a small calibration set, without constructing contrast samples or extra supervision. During inference, we leverage the resulting hallucination-related subspace to perform an online projection-based intervention on intermediate hidden states to suppress the hallucination-related directions, mitigating hallucinations while preserving language quality. Our RIVS is training-free and computationally efficient. Experiments on four hallucination and two reasoning datasets demonstrate that RIVS consistently reduces hallucinations in both long and short sequence generation tasks.

1 Introduction

In recent years, Large Vision-Language Models (LVLMs) [Liu *et al.*, 2023; Zhu *et al.*, 2023; Dai *et al.*, 2023; Ye *et al.*, 2023; Bai *et al.*, 2023; Zhang *et al.*, 2023; Chen *et al.*, 2023; Liu *et al.*, 2024b; Dong *et al.*, 2024; Bai *et al.*, 2025] have achieved remarkable success in various downstream tasks like image understanding or multimodal reasoning, by aligning visual encoders with large language models. Despite these advances, object hallucination persists in LVLMs, which often generate non-existent objects, attributes, or relations, especially in open-ended scenarios, thus significantly undermining LVLMs reliability in practice.

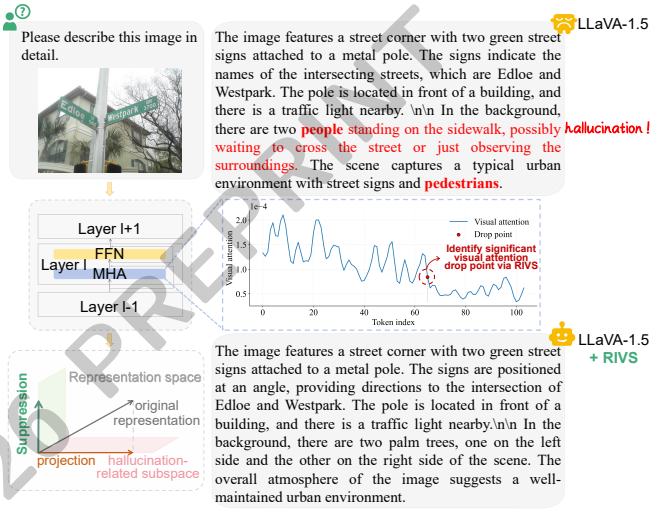

Figure 1: Hallucinations in LVLMs are often accompanied by decay in visual attention during generation. RIVS identifies the visual attention significant drop point, and constructs a hallucination-related subspace based on representation differences around it. During inference, suppressing the projection of hidden states onto this subspace mitigates hallucinations while preserving generation quality.

To address this issue, existing researches can be broadly categorized into three classes. Training-based methods often rely on data augmentation or fine-tuning, such as reconstructing data distributions [Liu *et al.*, 2024a], introducing counterfactual samples [Yu *et al.*, 2024a], or designing contrastive learning objective [Jiang *et al.*, 2024], which incur substantial computation cost and limits scalability. Inference-based methods, including contrastive decoding [Leng *et al.*, 2024; Chen *et al.*, 2024; Huo *et al.*, 2025; Zhuang *et al.*, 2025] , over-trust penalty [Huang *et al.*, 2024], and attention reweighting [Zhang *et al.*, 2024; Liu *et al.*, 2024c; Yin *et al.*, 2025], offering a lighter and more efficient solution, but provide limited interpretability of underlying model mechanisms. Post-hoc approaches commonly refine outputs with extra models [Yin *et al.*, 2024] or hallucination correctors [Zhou *et al.*, 2024], inevitably increasing computational complexity. Recently, some works begin to conduct transparency research on models through representation engineer-

*Corresponding author.

ing [Zou *et al.*, 2023], typically deriving steering vectors by computing averaged differences between positive and negative samples, relying on prior information [Khayatan *et al.*, 2025], carefully designed contrastive data [Liu *et al.*, 2025], or removing solely at specific individual objects [Jiang *et al.*, 2025]. Although effective under predefined conditions, above methods offer limited insight into how hallucinations emerge from the model's internal representations during the auto-regressive generation.

To address this limitation, we study hallucination from the perspective of representation shift induced by the model itself, focusing on how visual grounding representations change during auto-regressive generation. By analyzing attention patterns in intermediate layers, we observe that visual attention gradually decays as generation progresses, and hallucinations tend to emerge shortly after a pronounced drop in visual attention.

Motivated by this observation, we propose a two-stage hallucination mitigation framework, termed **R**epresentation **I**ntervention based on **V**isual Grounding **S**hift (**RIVS**). In the representation shift identification stage, we use a small calibration set to analyse visual attention dynamics and identify the generation step at which visual attention exhibits a significant decline for each sample. As shown in Fig. 1, such drop point is closely aligned with the generation of hallucinated token 'people', suggesting that hallucinations are not only triggered by individual token, but are associated with an internal representation shift from visual grounded to ungrounded. Using the identified drop point as an anchor, we extract difference vectors between hidden states before and after the decline to capture the direction of representation drift. By aggregating such vectors across samples and layers, we construct a low-dimensional hallucination-related subspace, without relying on explicit contrast sample construction or external supervision. In the representation intervention stage, we apply a lightweight and controllable online intervention during inference by projecting intermediate hidden states away from the constructed subspace. Our approach requires no model retraining or parameter modification, introducing minimal computational overhead.

The effectiveness of RIVS is evaluated on LLaVA-1.5, InstructBLIP, Qwen-VL and Qwen2.5-VL across six benchmarks: CHAIR [Rohrbach *et al.*, 2018], POPE [Li *et al.*, 2023], AMBER [Wang *et al.*, 2023], MME [Fu *et al.*, 2023], ScienceQA [Lu *et al.*, 2022] and MM-Vet [Yu *et al.*, 2024b]. Extensive experimental results validate the effectiveness of our RIVS in mitigating various types of hallucinations in LVLMs while preserving visual perception and reasoning capabilities.

To sum up, our main contributions are as follows:

- We present an analysis of hallucination in LVLMs from the perspective of internal representation shift during auto-regressive generation, demonstrating the shifts induced by visual attention decay can provide valuable guidance for mitigating hallucinations.
- We propose RIVS, a novel two-stage framework that identifies hallucination-related representation shifts using a small calibration set and constructs a low-dimensional hallucination subspace without explicit contrastive sample construction, enabling effective intervention during inference.
- Extensive experiments on six benchmarks show that RIVS consistently reduces hallucinations across multiple LVLMs, while preserving language quality and achieving favorable inference efficiency.

2 Related Work

Object hallucinations in LVLMs refer to models erroneously generating content that does not correspond to the input image. Fine-tuning methods aim to balance data distributions [Liu *et al.*, 2024a], construct counterfactual data [Yu *et al.*, 2024a], and modify optimisation objectives [Jiang *et al.*, 2024] for fine-tuning LVLMs. Inference-time approaches intervene during decoding stage. VCD [Leng *et al.*, 2024] contrasts hallucinations induced from language priors by introducing noise to images. SID [Huo *et al.*, 2025] amplifies multi-modal hallucinations by retaining unimportant visual tokens, then eliminates them through contrastive decoding. HALC [Chen *et al.*, 2024] mitigates hallucinations by focusing on the most critical image regions. VASparse [Zhuang *et al.*, 2025] designs a token selection strategy based on visual perception and introduces a sparse visual contrast decoding method to calibrate the output distribution. OPERA [Huang *et al.*, 2024] penalises generated tokens with attention sinking patterns, thereby adjusting the candidate level of decoded tokens. VAF [Yin *et al.*, 2025] enhances attention to the visual tokens in the model's intermediate layers, thereby strengthening the model's visual representations. Post-hocs methods [Yin *et al.*, 2024; Zhou *et al.*, 2024; Lee *et al.*, 2024] employ auxiliary models or train specialized hallucination correctors to rectify generated hallucinatory text.

Inspired by Representation Engineering [Zou *et al.*, 2023], recent work [Khayatan *et al.*, 2025] indicates behavioural changes in MLLMs can be characterised as representation shifts in the latent space. VTI [Liu *et al.*, 2025] constructs positive and negative samples to estimate intervention directions and manipulate latent representations in visual and textual modules. Unlike aforementioned methods, we re-examine the mechanism of hallucination generation in LVLMs from the perspective of representation shift during auto-regressive generation. This design avoids additional supervision costs and minimises disruption to language generation capabilities.

3 Preliminary

LVLMs takes an image V and a text prompt T as input, the image is first encoded by a pre-trained visual encoder, then projected into the semantic space of LLM via a connector to yield a sequence of visual features $X_v \in \mathbb{R}^{n\times d}$, where n denotes the number of visual tokens. The text input is tokenized into a sequence of language features $X_l \in \mathbb{R}^{m\times d}$, with m text tokens. The final multimodal input is formed by concatenating visual and language sequence as $X_{vl} = (X_v; X_l) \in \mathbb{R}^{(n+m)\times d}$.

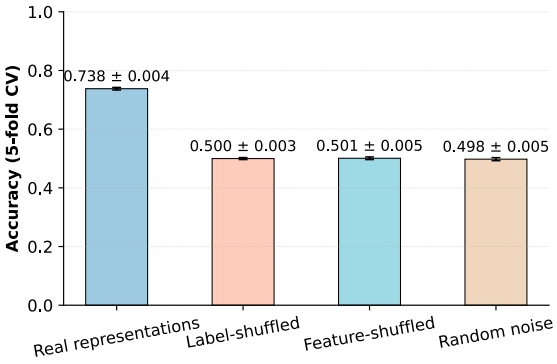

Figure 2: Comparison of Linear Probe Performances.

The decoder consists of L stacked Transformer blocks and generates output sequence $\mathbf{y} = [y_1, y_2, \ldots, y_T]$ autoregressively. Denote $H^{(l)}$ as the hidden states at layer l, each layer transforms the representations from the previous layer as:

$$H^{(l)} = Transformer(H^{(l-1)}), l = 1, 2, \ldots, L \quad (1)$$

4 Method

In this section, we first investigate the correlation between intermediate-layer representation shifts and hallucination generation in Sec. 4.1. Subsequently, in Sec. 4.2, we then introduce a lightweight representation intervention mechanism to mitigate hallucinations during inference.

4.1 Correlation Between Intermediate Representation Shifts and Hallucinations

Linear Artificial Tomography (LAT)

Linear Artificial Tomography (LAT) [Zou *et al.*, 2023] is a technique in Representation Engineering for analysing and manipulating high-level concepts encoded in neural networks, which comprises two components: *representation reading* and *representation control*. Representation reading locates representations of high-level concepts within the network by: (1) designing stimulus and task, (2) collecting neural activity, and (3) constructting a linear model. Representation control subsequently leverages extracted reading vectors or contrast vectors to intervene in model activations through linear operations such as projection or linear combination.

Hallucination Analysis via Visual Attention-Guided Difference Vectors

Prior studies suggest that hallucinations in LVLMs are closely associated with forgetting or weakening of visual grounding during generation [Deng *et al.*, 2024; Zou *et al.*, 2024]. Instead of explicitly construct contrast inputs [Liu *et al.*, 2025], we analyse hallucination generation by identifying *significant visual attention decay points* during generation and extracting representation difference vectors around these points.

Stimulus and Task. We consider an image captioning task, where LVLM is prompted with an image and a instruction (*e.g.*, "Please describe this image in detail"). During generation, we track the model's attention allocation to visual tokens and identify generation step at which visual attention exhibits a significant decline.

Neural activity collection. To investigate the impact of representations on hallucinations, we identify points where visual attention significantly drops during generation, then extract difference vectors around these drop points to capture variations in representations. Through this process, we can capture the shift in the model's understanding of visual information during the generation process.

Adaptive threshold-based drop point detection. For each sample, we first detect visual drop points in generated tokens by tracking variations in visual attention. For each step t, we compute the average attention weight of the generated token towards visual tokens at layer l:

$$m_t^{(l)} = \frac{1}{|\mathcal{I}|} \sum_{h=1}^{H} \sum_{k \in \mathcal{I}} \mathbf{A}_t^{(l,h)}[y_t, k], \quad \mathcal{I} = \{i \mid i \in [v_s, v_e]\}, \quad (2)$$

where $\mathbf{A}^{(l,h)}$ denotes the attention weight matrix of the h-th head at layer l, and $\mathcal{I}$ indexes visual tokens. We concatenate over generation steps to obtain the visual attention trajectory, then normalize each visual attention curve as:

$$m^{(l)} = [m_1^{(l)}, m_2^{(l)}, \ldots, m_T^{(l)}], \quad \tilde{m}^{(l)} = \frac{m^{(l)} - \mu^{(l)}}{\sigma^{(l)} + \epsilon}, \quad (3)$$

where $\mu^{(l)}$ and $\sigma^{(l)}$ denote mean and standard deviation. We further apply Gaussian smoothing $\mathcal{G}_\sigma(\cdot)$ to suppress noise:

$$\hat{m}^{(l)} = \mathcal{G}_\sigma(\tilde{m}^{(l)}), \quad (4)$$

Next, we define the first-order discrete derivative of the smoothed attention curve as:

$$\Delta^{(l)}(t) = \hat{m}_t^{(l)} - \hat{m}_{t-1}^{(l)}, \quad t = 2, \ldots, T \quad (5)$$

where $\Delta^{(l)}(t) < 0$ indicates a declining trend in visual attention at generation step t. We define 'significant drop' as a marked negative deviation relative to the overall variation distribution of this curve. Specifically, we propose a drop-point detection method based on adaptive threshold $\tau^{(l)}$, defined as the γ-quantile (usually set to 0.1) of the derivative distribution, and dentify the earliest generation step satisfying the derivative value falls below the threshold, as the significant visual attention drop point $t_\star^{(l)}$:

$$\begin{aligned} \tau^{(l)} &= \text{Quantile}_\gamma(\{\Delta^{(l)}(t)\}_{t=2}^{T}), \\ t_\star^{(l)} &= min\{t \mid \Delta^{(l)}(t) \leq \tau^{(l)}\}, \end{aligned} \quad (6)$$

If no such point exists, we select the step with the minimum derivative value as the most notable turning point:

$$t_\star^{(l)} = \arg\min_t \Delta^{(l)}(t), \quad (7)$$

Difference vector Extraction. To characterise the representation shift induced by the visual grounding decay, we extract representations from windows before and after the drop point $t_\star^{(l)}$, using them to derive difference vectors. The window preceding the visual drop point is denoted as $\mathcal{B} = \{t_\star^{(l)} - w, \ldots, t_\star^{(l)} - 1\}$, and the window after the visual drop point as $\mathcal{A} = \{t_\star^{(l)}, \ldots, t_\star^{(l)} + w - 1\}$. We then average representations within each window as:

$$\bar{\mathbf{h}}_{\text{pre}}^{(l)} = \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \mathbf{h}_t^{(l)}, \quad \bar{\mathbf{h}}_{\text{post}}^{(l)} = \frac{1}{|\mathcal{A}|} \sum_{t \in \mathcal{A}} \mathbf{h}_t^{(l)}, \quad (8)$$

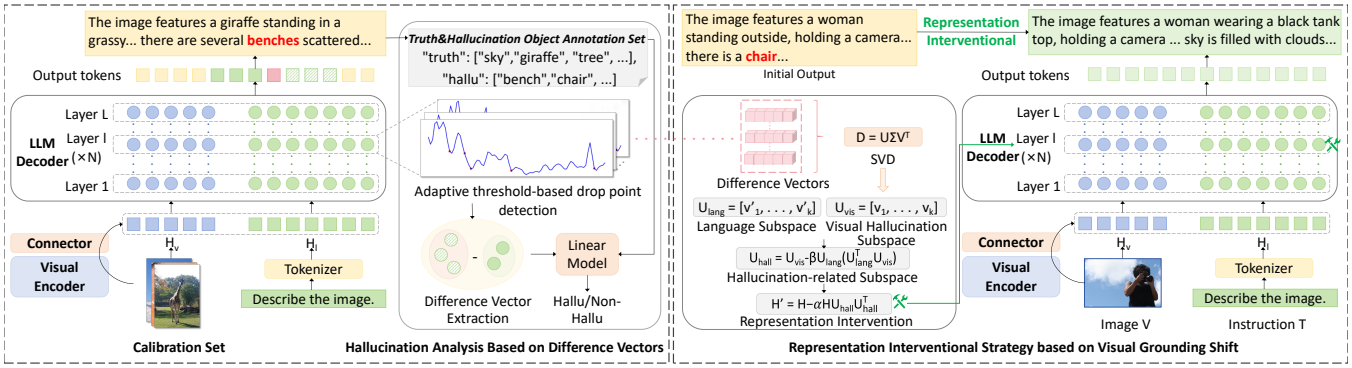

Figure 3: Overview of RIVS. Left: The process of extracting difference vectors based on visual attention shift during generation on the calibration set, following analysis of their association with hallucinations via hallucination annotations. Right: The process of constructing hallucination-related subspaces from difference vectors, subsequently employed to intervene intermediate representations during inference.

where $\mathbf{h}_t^{(l)}$ denotes the representation corresponding to the token generated at step t of layer l.

Next, the representation shift direction induced by visual grounding instability is then defined as:

$$\mathbf{d}^{(l)} = \frac{\bar{\mathbf{h}}_{\text{post}}^{(l)} - \bar{\mathbf{h}}_{\text{pre}}^{(l)}}{\left\| \bar{\mathbf{h}}_{\text{post}}^{(l)} - \bar{\mathbf{h}}_{\text{pre}}^{(l)} \right\|_2}, \tag{9}$$

Constructing a Linear Model. The objective of this step is to examine whether hallucination-related information is linearly encoded in the representation shift directions extracted from the model, using only internal neural activations as input. Following the representation reading paradigm of LAT, we conduct a linear probing experiment on the MSCOCO dataset [Lin *et al.*, 2014]. Specifically, we construct a *token-anchored representation shift signal* based on visual attention decay. Specifically, for each generated sequence $s_i = \{y_{i,1}, y_{i,2}, \ldots, y_{i,T}\}$, we identify the visual attention drop point $t_{i,\star}$ following the process described in neural activity collection. Using this position as an anchor, we extract representations from its adjacent windows to compute the difference vector d_i, which characterises the representation shift induced by visual grounding decay. For hallucination annotation, we map word-level annotations to token-level and assign each difference vector a binary label l_i based on whether hallucination occurs following its anchor point. After class-balanced sampling, we ultimately obtain a labelled dataset of difference vectors: $\mathcal{D} = \{(d_i, l_i)\}$.

To assess the linear separability of hallucination-related signals, we perform a linear probing experiment using difference vectors extracted from decoder layer 25. Specifically, we train a logistic regression classifier to predict hallucination occurrence based solely on difference vectors and evaluated with 5-fold cross-validation. The probe achieves an average AUC of 81.67 and accuracy of 73.79, indicating that hallucination-raleted information is linearly separable from the shifting representation captured by visual attention-guided difference vectors. Fig. 2 compares probes trained on difference vectors with several control baselines, including label shuffling, feature shuffling, and random noise. Probes trained on representation differences significantly outperform other baselines, with all pairwise comparisons showing statistically significant improvements ($p<0.001$), indicating the robustness of the observed linear separability.

4.2 Representation Intervention Strategy based on Visual Grounding Shift

Motivated by the linear separability analysis in the previous section, we observe that representation difference vectors before and after the visual attention drop point are strongly correlated with hallucination occurrence in the latent space. Building on this observation and inspired by representation control in LAT, we propose a representation intervention strategy based on visual grounding shift, as illustrated in Fig. 3. This approach employs a two-stage framework to decouple identification of hallucination-related directions from online intervention. First, we extract and aggregate representation difference vectors associated with visual attention decay to construct a low-dimensional hallucination subspace. In the inference stage, we apply a lightweight projection-based intervention to intermediate hidden states, suppressing representation shifts along hallucination-related directions during generation, thereby reducing hallucination occurrence.

Construction of the Hallucination Subspace

Considering that hallucinations in LVLMs are not arise from representations at a single layer, but rather result from coordinated shifts across multiple layers, we collect representation difference vectors across a set of layers $\mathcal{L} = \{l_1, l_2, \ldots, l_c\}$. Specifically, for each sample i and layer $\in \mathcal{L}$, we first track the visual attention curve $\tilde{m}^{(l)}$ over the generated sequence and detect the visual attention drop point $t_{i,\star}^{(l)}$ using the adaptive threshold-based drop point detection method described in Sec. 4.1, and then extract hidden states from windows before and after the drop point to obtain the difference vector $\mathbf{d}_i^{(l)}$.

To obtain a stable hallucination-related respresentation, we aggregate difference vectors across N samples in calibration set and $|\mathcal{L}|$ layers, all normalized vectors are concatenated as:

$$\mathbf{D} = \left[\mathbf{d}_1^{(l_1)} \quad \mathbf{d}_2^{(l_1)} \quad \cdots \quad \mathbf{d}_N^{(l_c)} \right]^\top \in \mathbb{R}^{M \times d}, \tag{10}$$

where $M = N \times |\mathcal{L}|$ denotes the number of direction vectors.

We then centre each difference vector as:

$$\mu_D = \frac{1}{M}\sum_{m=1}^{M}\mathbf{d}_m, \tag{11}$$

$$\tilde{\mathbf{d}}_m = \mathbf{d}_m - \mu_D, \quad m = 1, \ldots, M, \tag{12}$$

where $\mathbf{d}_m$ denote the normalized difference vector. Then, we apply principal component analysis (PCA) to extract the most representative low-dimensional subspace. Specifically, we perform singular value decomposition (SVD) on $\tilde{\mathbf{D}}$, which denote the centred version of $\mathbf{D}$, and select the top-k right singular vectors as an orthogonal basis:

$$\tilde{\mathbf{D}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \quad \mathbf{U}_{vis} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k], \tag{13}$$

This subspace characterises representation shift patterns associated with visual attention decay across layers and samples.

Language Subspace Modelling and Orthogonal Constraint

Since vision and language tokens in LVLMs are mapped into a shared semantic space, directly intervening on hidden representations may inadvertently undermine language generation. To mitigate this effect, we model a language-dominant subspace and constrain the hallucination space accordingly. Specifically, we collect the hidden states from $\mathcal{L}$ based solely on the text input, then perform PCA on these linguistic representations in the same manner as for the difference vectors, yielding a low-dimensional language subspace $\mathbf{U}_{lang}$, which captures the principal language-driven direction during generation.

To ensure the intervention primarily affects representations related to visual grounding without disrupting language semantic structures, we orthogonalise the visual hallucination subspace with respect to the language subspace:

$$\mathbf{U}_{hall} = \mathbf{U}_{vis} - \beta\mathbf{U}_{lang}(\mathbf{U}_{lang}^\top\mathbf{U}_{vis}), \tag{14}$$

where $\beta \in [0, 1]$ controls the strength of the orthogonal constraint. After normalization, $\mathbf{U}_{hall}$ serves as the final hallucination-related subspace, capturing representation shifts induced by visual grounding degradation while preserving natural language structures.

Projection-based Representation Intervention

After constructing the hallucination-related subspace, we employ it for online representation intervention during inference to suppress hallucination-related representation shifts. Unlike methods based on retraining or modifying model weights, we adopt a lightweight linear projection mechanism to modify the model's intermediate representations. For the hidden states $H^{(l)}$ at layer l, we apply the intervention as:

$$\begin{aligned} H'^{(l)} &= H^{(l)} - \alpha \cdot P_{\mathbf{U}_{hall}}(H^{(l)}), \\ P_{\mathbf{U}_{hall}}(H^{(l)}) &= H^{(l)}\mathbf{U}_{hall}\mathbf{U}_{hall}^\top, \end{aligned} \tag{15}$$

where α controls the intervention strength. As hallucination-related representation shifts emerge across multiple layers, the intervention is applied to the hidden states of all layers $l \in \mathcal{L}$ respectively.

5 Experiments

5.1 Experimental Settings

Models. Following previous works [Liu *et al.*, 2025; Yin *et al.*, 2025], we conduct experiments on four mainstream models: LLaVA-1.5 [Liu *et al.*, 2024b], Instructblip [Dai *et al.*, 2023], Qwen-VL [Bai *et al.*, 2023] and Qwen2.5-VL [Bai *et al.*, 2025]. All model are with 7B parameters.

Baselines. We compare performance with basic greedy search and several popular training-free methods, including *VCD* [Leng *et al.*, 2024], *OPERA* [Huang *et al.*, 2024], *HALC* [Chen *et al.*, 2024], *SID* [Huo *et al.*, 2025], *VTI* [Liu *et al.*, 2025], *VASparse* [Zhuang *et al.*, 2025] and *VAF* [Yin *et al.*, 2025]. To ensure a fair comparison, we use the paper's default settings for these baselines.

Implementation Details. The sample size N for calibration set is set to 100. We uniformly set $k = 3$, $\alpha = 0.1$, and set β to 0.5, 0.8, 0.2 and 1.0, w to 100, 6, 100, 100 for LLaVA-1.5, InstructBLIP, Qwen-VL, and Qwen2.5-VL respectively. We faithfully replicate all baselines using their released code and the hyperparameters reported in the original papers.

5.2 Experimental Results

We evaluate on four hallucination benchmarks and two reasoning benchmarks.

Hallucination Mitigation

CHAIR. [Rohrbach *et al.*, 2018] We randomly select 500 images from MSCOCO val2014 [Lin *et al.*, 2014], and use the prompt "`Please describe this image in detail.`" to query LVLMs. To ensure fair comparison, we set the max new tokens to 512. Table 1 demonstrates that our method outperforms all baselines except VTI [Liu *et al.*, 2025] in mitigating hallucinations on all LVLMs. While VTI yields a low hallucination score, it produces shorter outputs with noticeably reduced recall, which limit ground-truth object coverage. In contrast, our RIVS achieves a more favorable trade-off between factuality and content coverage.

POPE. [Li *et al.*, 2023] We evaluate our method on MSCOCO using three settings: *random*, *popular*, and *adversarial*. We report the average performance across the three settings in Table 2, indicating that our approach achieves superior average performance across all models compared to the baselines, with particularly enhanced capabilities in the more challenging popular and adversarial settings.

AMBER. [Wang *et al.*, 2023] AMBER is a benchmark for comprehensively evaluating discriminative and generative capabilities of LVLMs, and the AMBER Score provides a synthetic metric for this assessment. Similarly, as shown in Table 6, while VTI effectively eliminates hallucinations in the generation task, its coverage rate is significantly reduced compared to the original algorithm (*40.0 vs 50.5*). In contrast, our approach not only mitigates hallucinations effectively but also maintains coverage rate, further demonstrating our method's capability to reduce hallucinations while preserving content richness. Furthermore, our approach achieves the best overall performance, demonstrating that pre-extracting difference vectors from a small sample set is task-agnostic and performs excellently on both short and long sequence

Method	LLaVA-1.5				InstructBLIP			
	$C_S\downarrow$	$C_I\downarrow$	Recall↑	Length	$C_S\downarrow$	$C_I\downarrow$	Recall↑	Length
Greedy	51.6	13.6	77.6	<u>101.0</u>	50.3	13.5	71.4	101.5
$VCD_{\text{CVPR'24}}$	56.8	16.1	78.1	**104.7**	58.0	15.8	68.6	**111.6**
$OPERA_{\text{CVPR'24}}$	46.3	12.9	<u>78.2</u>	95.7	49.8	14.3	**73.5**	97.3
$HALC_{\text{ICML'24}}$	42.0	12.9	71.7	96.5	61.4	19.1	71.8	<u>104.0</u>
$SID_{\text{ICLR'25}}$	44.2	11.6	**78.4**	95.1	46.6	<u>12.7</u>	71.7	96.7
$VTI_{\text{ICLR'25}}$	**12.0**	**5.0**	54.5	38.9	-	-	-	-
$VAF_{\text{CVPR'25}}$	49.8	13.3	77.9	99.4	<u>44.0</u>	13.0	<u>72.1</u>	98.8
$VASparse_{\text{CVPR'25}}$	51.0	13.4	76.4	97.9	-	-	-	-
RIVS (Ours)	<u>39.2</u>	<u>10.3</u>	76.8	98.5	**34.0**	**11.1**	70.5	85.9

Table 1: Results on CHAIR. The best performances are **bolded**, the second-best results are underscore. '-' indicates that the implementation of the corresponding method on this model is not open-source or no results have been published for it.

Method	LLaVA-1.5		InstructBLIP	
	Acc.↑	F1↑	Acc.↑	F1↑
Greedy	84.0	83.6	81.4	81.6
$VCD_{\text{CVPR'24}}$	84.2	84.1	81.8	81.3
$OPERA_{\text{CVPR'24}}$	85.4	83.9	84.9	84.8
$HALC_{\text{ICML'24}}$	83.5	84.1	82.1	83.4
$SID_{\text{ICLR'25}}$	85.2	85.4	83.7	84.1
$VTI_{\text{ICLR'25}}$	<u>86.0</u>	85.6	-	-
$VAF_{\text{CVPR'25}}$	86.0	85.0	84.6	83.9
$VASparse_{\text{CVPR'25}}$	85.8	<u>85.7</u>	-	-
RIVS (Ours)	**86.8**	**86.5**	**85.4**	**85.3**

Table 2: Results on POPE. Since the implementations of the VASparse and VTI on InstructBLIP are not publicly available, results are shown only on LLaVA-1.5, same as for other benchmarks.

Method	LLaVA-1.5		InstructBLIP	
	Hallu.↑	Total↑	Hallu.↑	Total↑
Greedy	621.6	1452.7	415.0	1120.1
$VCD_{\text{CVPR'24}}$	593.3	1314.9	415.0	1122.2
$OPERA_{\text{CVPR'24}}$	626.7	1448.6	**425.0**	1121.2
$HALC_{\text{ICML'24}}$	621.7	1444.6	423.3	934.5
$SID_{\text{ICLR'25}}$	<u>626.7</u>	1458.4	415.0	1121.8
$VTI_{\text{ICLR'25}}$	525.0	1213.4	-	-
$VAF_{\text{CVPR'25}}$	625.0	<u>1477.9</u>	420.0	<u>1137.9</u>
$VASparse_{\text{CVPR'25}}$	606.7	1397.1	-	-
RIVS (Ours)	**631.6**	**1500.9**	<u>423.3</u>	**1177.1**

Table 3: Results on all MME perception-related tasks.

tasks. Moreover, whereas other approaches demonstrate effectiveness on LLaVA-1.5 yet exhibit performance degradation or marginal gains on InstructBLIP, our method consistently achieves superior performance across diverse LVLMs.

MME. [Fu *et al.*, 2023] Table 3 presents performance across ten perception tasks, where Hallu. denotes scores on hallucination subset. Our approach obtains considerable performance on hallucination subset and achieves the best overall performance. Specifically, it surpasses the optimal baseline by 22 and 39 points on LLaVA-1.5 and InstructBLIP respectively, demonstrating the superiority of the proposed method in visual perception.

Furthermore, we validate the effectiveness of our approach on the Qwen-VL and Qwen2.5-VL model, which possesses stronger multimodal capabilities. The results in Table 4 demonstrate that RIVS consistently achieves performance gains even on the less-hallucinatory LVLM.

Setting		POPE		CHAIR		AMBER
Model	Method	Acc.↑	F1↑	$C_S\downarrow$	$C_I\downarrow$	Score↑
Qwen-VL	Greedy	84.7	83.5	15.4	8.3	91.0
	RIVS (Ours)	**87.5**	**87.0**	**12.5**	**7.6**	**92.1**
Qwen2.5-VL	Greedy	86.5	85.5	31.2	8.7	90.7
	RIVS (Ours)	**87.5**	**86.2**	**25.2**	**7.5**	**92.0**

Table 4: Results of Qwen series models Qwen-VL and Qwen2.5-VL. Since other baselines have not been publicly implemented on this model, comparisons are only with vanilla decoding.

Benchmark	Greedy	VCD	VTI	**RIVS (Ours)**
ScienceQA (Accuracy)	64.3	60.0	61.8	**64.9**
MM-Vet (GPT-4 Score)	31.5	28.8	28.1	**32.8**

Table 5: Results of LLaVA-1.5 on complex reasoning benchmarks.

Comprehensive Visual-Language Abilities Assessment

To evaluate our method's capability in vision-language reasoning, we further experiment on ScienceQA [Lu *et al.*, 2022] and MM-Vet [Yu *et al.*, 2024b]. The results in Table 5 demonstrate that while VCD and VTI mitigates hallucinations to some extent, it compromises model's reasoning ability. By contrast, our approach preserves reasoning and response accuracy, proving better adapted to complex reasoning tasks.

Inference efficiency

Fig.4 compares the inference efficiency between different methods. To ensure a fair comparison, inference latency was measured in a token-by-token manner. Compared to other methods, our approach achieves the best performance in inference speed and also maintains a significant advantage in GPU memory usage. The bubble area is proportional to F1 Score on POPE, and the result indicates that our method achieves an optimal balance between performance and computational cost overall.

Model	Method	GENERARIVE TASK				DISCRIMINATIVE TASK				AMBER Score↑
		CHAIR↓	*Cover*↑	*Hal*↓	*Cog*↓	*Acc.*↑	*P.*↑	*R.*↑	*F1*↑	
LLaVA-1.5	Greedy	7.2	50.5	32.7	3.9	72.1	91.8	62.9	74.6	83.7
	$VCD_{\text{CVPR'24}}$	11.0	50.9	45.7	3.9	77.0	87.1	76.6	81.5	85.2
	$OPERA_{\text{CVPR'24}}$	7.4	49.0	36.8	3.8	76.6	92.2	70.2	79.7	86.1
	$HALC_{\text{ICML'24}}$	6.2	48.7	29.4	2.6	72.4	80.2	78.5	79.3	86.5
	$SID_{\text{ICLR'25}}$	5.5	**52.3**	27.5	2.4	75.0	**93.7**	66.1	77.5	88.0
	$VTI_{\text{ICLR'25}}$	**3.4**	40.0	**11.3**	**1.0**	76.3	84.2	78.6	81.3	88.9
	$VASparse_{\text{CVPR'25}}$	6.3	50.1	28.3	3.3	77.0	92.0	71.0	80.1	86.9
	$VAF_{\text{CVPR'25}}$	6.9	50.5	32.8	3.6	79.2	90.6	76.5	83.0	88.0
	RIVS (Ours)	6.1	51.3	31.1	3.1	**81.0**	91.3	**78.9**	**84.6**	**89.2**
InstructBLIP	Greedy	7.3	53.8	35.0	3.7	76.3	81.2	**80.0**	80.6	86.6
	$VCD_{\text{CVPR'24}}$	10.1	54.1	46.9	4.8	73.4	87.5	69.9	77.7	83.8
	$OPERA_{\text{CVPR'24}}$	8.0	51.0	35.0	3.8	74.5	83.9	75.7	79.6	85.8
	$HALC_{\text{ICML'24}}$	14.0	**55.2**	57.0	6.7	75.3	**94.6**	67.1	78.5	82.2
	$SID_{\text{ICLR'25}}$	7.5	53.6	35.5	3.9	75.9	83.2	79.5	81.3	86.9
	$VAF_{\text{CVPR'25}}$	8.2	54.4	37.1	4.3	77.8	88.8	76.0	81.9	86.8
	RIVS (Ours)	**6.2**	53.3	**29.3**	**3.0**	**79.0**	89.8	77.1	**83.0**	**88.4**

Table 6: Results on AMBER. The best performances are **bolded**.

Intervention Strategy	C_S↓	C_I↓	Recall↑	Length
LLaVA-1.5	51.6	13.6	77.6	101.0
Random Difference Vectors	50.6	13.5	77.6	100.6
Random Subspace	52.6	13.9	78.2	100.5
Random Position	43.8	11.5	75.3	97.6
Visual Subspace	44.4	11.9	74.6	96.6
Projection Amplification	57.8	15.6	**80.2**	**108.5**
RIVS	**39.2**	**10.3**	76.8	98.5

Table 7: Ablation study of intervention strategies on MSCOCO.

Ablation Study

To validate the effectiveness of the proposed method, we conduct ablation studies with several controlled baselines, including: (1) randomize difference vectors, (2) hallucination-related subspace constructed from randomly sampled orthogonal bases, (3) difference vectors extracted from randomly selected positions, (4) employing the visual hallucination-releated subspace directly for projection intervention without applying language orthogonalisation, and (5) projection amplification, where the projection onto the hallucination-related subspace is added rather than suppressed.

As shown in Table 7, random difference vectors yield limited improvements, indicating that hallucination mitigation cannot be achieved by arbitrary representation interventions. Random position and visual hallucination-related subspace reduce hallucination metrics to some extent, indicating that suppressing representation drift can partially alleviate hallucinations, but at the cost of modest decreases in recall and generation length. This indicates that inadequate precision in suppressing representations may lightly impair model's language generation capability. In contrast, random subspace and projection amplification substantially increase hallucination metrics, confirming that the identified subspace indeed encodes hallucination-related informations. Our RIVS achieves the lowest hallucination scores while maintaining comparable Recall and generation length, demonstrating effective hallucination suppression without sacrificing generation quality.

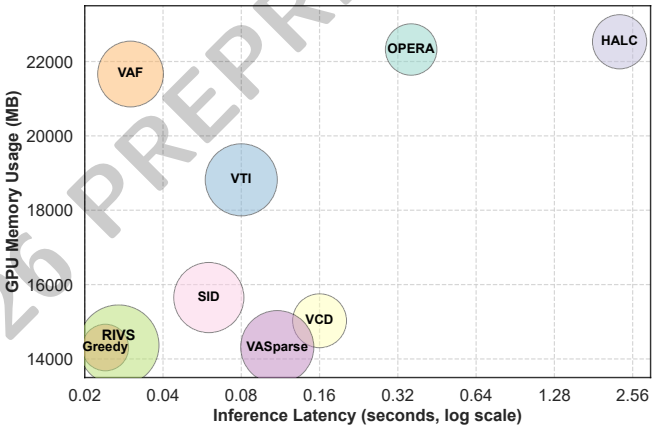

Figure 4: Comparison of computational efficiency of different methods used for hallucination mitigation using LLaVA-1.5 (max new tokens is 512). The area of each bubble represents the F1 score on POPE, larger areas indicate better performance.

6 Conclusion

In this work, we investigate object hallucination in LVLMs from the perspective of dynamic representation shift during auto-regressive generation. By analyzing the decay of visual attention, we identify a shift in representations from visually grounded to language-dominated states, which is closely associated with hallucination. Building on this observation, we propose RIVS, a training-free representation intervention framework that constructs a hallucination-related subspace from representation differences around visual attention drop points and applies an projection-based intervention during inference. Extensive experiments across multiple hallucination and reasoning benchmarks demonstrate that RIVS effectively reduces hallucinations while preserving language quality and generation richness, offering a new representational perspective for understanding and mitigating hallucinations in LVLMs.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62276110 and in part by the fund of Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL). The authors would also like to thank the anonymous reviewers for their comments on improving the quality of this paper.

References

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[Bai *et al.*, 2025] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[Chen *et al.*, 2023] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[Chen *et al.*, 2024] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

[Dai *et al.*, 2023] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

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

[Dong *et al.*, 2024] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*, 2024.

[Fu *et al.*, 2023] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

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

[Huo *et al.*, 2025] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

[Jiang *et al.*, 2024] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 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, 2024.

[Jiang *et al.*, 2025] Nick Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*, 2025.

[Khayatan *et al.*, 2025] Pegah Khayatan, Mustafa Shukor, Jayneel Parekh, Arnaud Dapogny, and Matthieu Cord. Analyzing finetuning representation shift for multimodal llms steering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2206–2216, 2025.

[Lee *et al.*, 2024] Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 391–404, 2024.

[Leng *et al.*, 2024] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

[Li *et al.*, 2023] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305, 2023.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in*

neural information processing systems, 36:34892–34916, 2023.

[Liu *et al.*, 2024a] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations,ICLR 2024*, 2024.

[Liu *et al.*, 2024b] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[Liu *et al.*, 2024c] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer, 2024.

[Liu *et al.*, 2025] Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Lu *et al.*, 2022] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

[Rohrbach *et al.*, 2018] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.

[Wang *et al.*, 2023] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

[Ye *et al.*, 2023] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

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

[Yin *et al.*, 2025] Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14625–14634, 2025.

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

[Yu *et al.*, 2024b] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, 2024.

[Zhang *et al.*, 2023] Pan Zhang, Xiaoyi Dong Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Hang Yan, et al. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023.

[Zhang *et al.*, 2024] Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms. *CoRR*, abs/2411.09968, 2024.

[Zhou *et al.*, 2024] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

[Zhu *et al.*, 2023] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[Zhuang *et al.*, 2025] Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. Vasparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4189–4199, 2025.

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

[Zou *et al.*, 2024] Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, et al. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*, 2024.