DYNAMIC MULTIMODAL ACTIVATION STEERING FOR HALLUCINATION MITIGATION IN LARGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models (LVLMs) exhibit outstanding performance on vision-language tasks but struggle with hallucination problems. Through in-depth analysis of LVLM activation patterns, we reveal two key findings: 1) truthfulness and visual perception capabilities predominantly engage different subsets of attention heads within the model architecture; and 2) truthfulness steering vectors vary significantly across different semantic contexts. Based on these observations, we propose Dynamic Multimodal Activation Steering, a training-free, plug-and-play approach for hallucination mitigation. Our method constructs a semantic-based truthfulness steering vector database and computes visual perception steering vectors, enabling context-aware interventions during inference by dynamically selecting the most relevant steering vectors based on input semantic similarity and applying them to the most influential attention heads. We conduct comprehensive experiments across multiple models and datasets, demonstrating that our approach significantly enhances model performance, outperforming existing state-of-the-art methods.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) have demonstrated remarkable performance on visual question answering (VQA), image captioning, and related tasks Liu et al. (2023; 2024c); Bai et al. (2023); Chen et al. (2024); Dai et al. (2023). However, these models suffer from significant hallucination phenomena, manifested as fabricating non-existent objects or incorrectly describing image content Liu et al. (2024b); Bai et al. (2024). Such hallucinations limit the applicability of LVLMs in downstream applications including autonomous driving Cui et al. (2024), robotics Li et al. (2024b), and other safety-critical domains.

Due to the complex architecture of LVLMs, the causes of multimodal hallucinations are diverse. To address these multimodal hallucination issues, numerous approaches have been proposed Leng et al. (2024); Huang et al. (2024); An et al. (2025); Yin et al. (2024); Liu et al. (2024a), which can be broadly categorized into two classes: training-based and decoding-based methods. Training-based methods primarily focus on constructing less biased datasets to fine-tune LVLMs, such as LRV Liu et al. (2024a), or employing reinforcement learning to train LVLMs, as demonstrated by RLHF-V Yu et al. (2024a). The limitations of these approaches lie in their requirements for carefully curated data and substantial computational resources, as well as the need to retrain models separately for different architectures. Decoding-based methods, on the other hand, modify the decoding strategies of LVLMs, such as VCD Leng et al. (2024) and ICD Wang et al. (2024b). While these methods avoid the need for training, they often compromise the quality of the generated content Yin et al. (2025).

More recently, researchers have begun investigating activation engineering Zou et al. (2023); Li et al. (2023b); Wang et al. (2024a) as an alternative approach to reduce hallucinations through targeted intervention in model representations. ICT Chen et al. (2025) is an image-object cross-level trusted intervention method that mitigates model hallucinations by applying noise to both images and objects, thereby enhancing the model's attention to visual information. However, this approach primarily focuses on visual-level interventions, neglecting the multimodal characteristics of LVLMs.

VTI Liu et al. (2025) intervenes in the hidden states of both the visual encoder and large language model during inference by pre-computing steering vectors for visual and textual modalities. Nevertheless, this method employs fixed steering vectors regardless of input variation, ignoring potential semantic differences across diverse contexts. The uniformly applied steering vectors fail to account for the nuanced semantic variations that exist across different inputs.

To address these challenges, we propose dynamic multimodal activation steering (DMAS), a training-free, plug-and-play approach for hallucination mitigation in LVLMs. Our method focuses on two types of attention heads in LVLMs: truthfulness-related and visual perception-related. For truthfulness heads, we explicitly model how truthfulness steering vectors vary across semantic contexts. We cluster data semantically and create sample pairs with and without hallucinations within each cluster. By contrasting attention activations between factual and hallucination-prone samples, we extract truthfulness steering vectors. These vectors are stored alongside their cluster embeddings in a key-value database.

For visual perception, we calculate activation differences between noise-free and noisy image inputs to derive perception steering vectors that enhance visual attention. During inference, we dynamically retrieve the most semantically relevant truthfulness steering vector for the input query and apply both truthfulness and visual perception vectors to the top-K attention heads with the largest activation differences. This dual intervention effectively reduces hallucinations. The main contributions of our paper are:

- We investigate activation differences in LVLMs, revealing that truthfulness and visual perception capabilities predominantly engage different subsets of attention heads, and demonstrate that truthfulness vectors vary significantly across different semantic contexts through visualization, indicating the necessity for dynamic rather than static intervention approaches.
- We propose dynamic multimodal activation steering, a training-free method for hallucination mitigation that constructs a semantic-based truthful steering vector database and visual perception steering vector, enabling context-aware interventions during inference by dynamically selecting appropriate steering vectors based on input semantic similarity.
- We conduct comprehensive experiments on multiple models across discriminative tasks and open-ended generation datasets. The experimental results demonstrate that our method achieves significant improvements: increasing total scores by 94.66 on MME and reducing 20.2% hallucinations on CHAIR, outperforming existing state-of-the-art methods. These results highlight the effectiveness of our approach in hallucination mitigation.

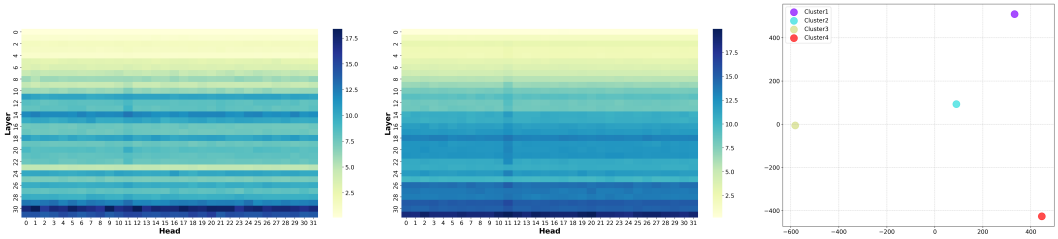
2 Related Work

2.1 Large Vision-Language Models

Large Vision-Language Models (LVLMs) have recently undergone rapid development, achieving excellent performance in image captioning and VQA tasks Yin et al. (2023); Jin et al. (2024). They typically consist of a vision encoder, a connection layer, and an LLM. As for the vision encoder, the VIT from CLIP Radford et al. (2021) is commonly used. For the connection layer, some models use simple MLP layers for alignment, such as LLaVA Liu et al. (2023; 2024c), Shikra Chen et al. (2023), PandaGPT Su et al. (2023), etc.; some models use Q-former for alignment, like BLIP2 Li et al. (2023a), InstructBLIP Dai et al. (2023), etc.; while others design special architectures. However, these LVLMs suffer from serious hallucination problems, and effectively eliminating hallucinations remains a popular research topic.

2.2 Hallucination Mitigation for LVLMs

Recently, numerous approaches have been proposed to mitigate multimodal hallucinations Liu et al. (2024b); Bai et al. (2024), addressing this issue across three key stages: training, inference, and post-processing. At the training stage, some research concentrates on constructing better data to train models. For example, LRV Liu et al. (2024a) constructs a high-quality instruction fine-tuning dataset containing balanced positive and negative samples, while other studies introduce reinforcement learning to the multimodal domain to reduce hallucination, such as RLHF-V Yu et al. (2024a) and RLAIF-V Yu et al. (2024b). These methods typically require carefully constructed training data



(a) The difference of truthfulness attention head activation.

(b) The difference of visual perception attention head activation.

(c) Steering vectors visualization.

Figure 1: Activation differences in LLaVAv1.5.

and consume substantial computational resources during training. Research on mitigating hallucinations at the inference stage often requires no training. VCD Leng et al. (2024) uses the distribution from noise-added images and the original output distribution to jointly determine the final distribution to mitigate hallucination. ICD Wang et al. (2024b) reduces hallucinations by contrasting output distributions between standard and deliberately disturbed instructions. These contrastive decoding methods often compromise the quality of the generated content Yin et al. (2025). Post-processing approaches correct the generated content from LVLMs to achieve hallucination reduction effects. For instance, LURE Zhou et al. (2024) constructs a dataset to train a hallucination revisor. However, these methods require the construction of a complex pipeline and increase the time required to obtain final outputs. To overcome these limitations, we propose dynamic multimodal activation steering, a training-free plug-and-play approach to mitigate hallucination in LVLMs by dynamically intervening in attention heads during inference time.

3 Preliminary Study

To understand the internal mechanisms underlying multimodal hallucinations, we conduct a systematic analysis of attention patterns in LLaVAv1.5 Liu et al. (2024c) across 3,000 samples from the SEED Li et al. (2024a) and AMBER Wang et al. (2023) datasets. Our investigation focuses on identifying which attention heads are most sensitive to truthfulness versus visual perception.

We design two complementary experiments to isolate attention mechanisms responsible for different aspects of multimodal processing. In the first experiment, we examine truthfulness related attention head by contrasting model activations when processing identical visual inputs paired with text prompts either with ground truth or hallucinated answers. This approach enables us to identify attention heads most relevant to truthfulness, we measure how each head's activation changes between truthful and hallucinated content by computing the difference: truthful activation minus hallucinated activation. In the second experiment, we investigate visual perception related attention head by comparing activations between clean images and their noise-corrupted counterparts, calculating activation differences by subtracting the activation values of non-noisy inputs from those with noise. As shown in Figures 1a and 1b, the activation patterns differ significantly between these two experiments. For truthfulness (Figure 1a), the most active attention heads appear predominantly in layer 30. In contrast, for visual perception (Figure 1b), the highest activation differences concentrate in layer 31. These distinct activation patterns provide a foundation for our targeted intervention approach that addresses both aspects simultaneously.

Furthermore, we divide the SEED and AMBER datasets into four semantic clusters and compute the activation differences for each cluster. Using t-SNE to visualize these differences in a two-dimensional space (Figure 1c), we observe a clear separation between clusters, with each occupying a distinct region in the projection space. This separation indicates that truthfulness direction vectors vary significantly across different semantic contexts. The heterogeneity in these patterns suggests that a static intervention approach would be insufficient, as it cannot account for the semantic-dependent nature of hallucinations. This observation directly motivates our dynamic multimodal activation steering method, which can adaptively select appropriate steering vectors based on the semantic content of the input query.

4 METHOD

In this section, we introduce dynamic multimodal activation steering. As shown in Figure 2, the method has three steps: the first step is to establish a dynamic truthfulness steering vector database, the second step is to calculate the steering vector for the model's visual perception attention heads, and the third step is to apply dynamic interventions to different attention heads during inference.

4.1 TRUTHFULNESS STEERING VECTOR DATABASE

We select the AMBER Wang et al. (2023) and SEED Li et al. (2024a) datasets as our data sources and divide the datasets into 4 clusters based on semantics. The questions in these two datasets are in the form of multiple-choice and discriminative questions, making it easy for us to create hallucinated answers for each sample (for discriminative questions, we change the answer to the opposite; for multiple-choice questions, we randomly select an incorrect option). Each cluster C_i comprises the question prompt T, visual input V, ground truth response Y_{pos}, and incorrect response Y_{neg} for every sample.

We input $(V, T + Y_{pos})$ and $(V, T + Y_{neg})$ separately into LVLMs and preserve the attention head activation values of the last token at each layer, denoted as A_{pos} and A_{neg}. We define the truthfulness steering vector as the activation difference between non-hallucinated outputs and hallucinated outputs within each cluster according to Equation 1:

$$D_i = \frac{1}{|C_i|} \sum_{j \in C_i} (A_{pos,j} - A_{neg,j}) \tag{1}$$

$|C_i|$ represents the number of samples in cluster C_i, and j indexes the samples within the cluster. Subsequently, we apply principal component analysis (PCA) to D_i to reduce insignificant noise, thereby extracting the principal components that influence truthfulness. The magnitude of D_i effectively quantifies the significance of each attention head in governing this specific model behavior.

Next, we construct a truthful steering vector database where the average embedding representation of questions from each cluster serves as the key, with the corresponding steering vector D_i as the value. During inference, our approach dynamically matches the semantic content of the input question to retrieve the most semantically similar steering vector, enabling context-appropriate interventions. We obtain key embeddings in the database and input text embeddings via sentence transformer.[1]

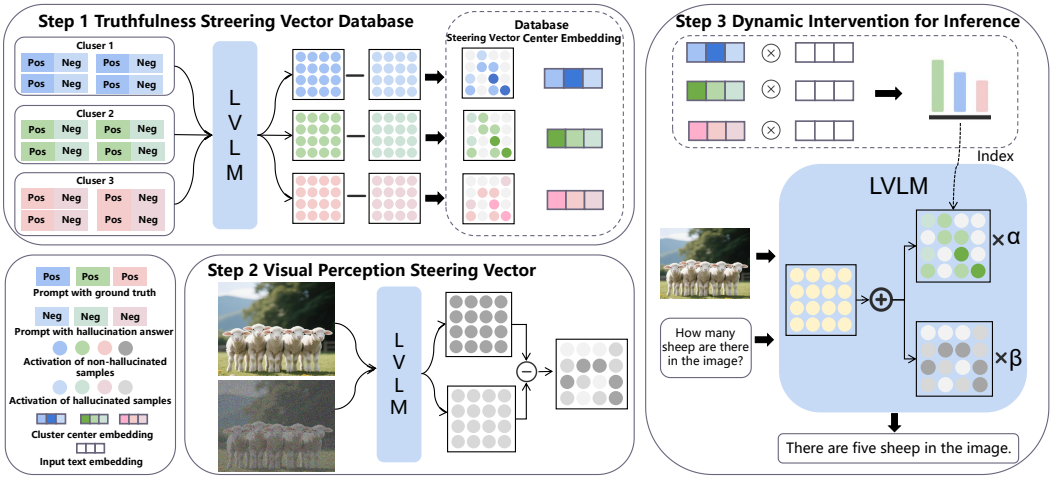


Figure 2: Overview of the DMAS framework.

[1] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

4.2 Visual Perception Steering Vector

Given a visual input V and a distorted visual input V' (obtained by adding noise to the image following the forward diffusion process in image generation Ho et al. (2020)), we first input V into an object detector YOLOv11 Khanam & Hussain (2024) to obtain a list of objects O present in the image, and insert them into a simple template 'The image depicts objects ' , denoted as Y_O. At the same time, we randomly select an equal number of objects O' from a predefined object library that are not in O, also inserting them into the template, denoted as $Y_{O'}$. The prompt T is fixed as 'Please describe this image.' Next, we obtain the final inputs $(V, T + Y_O)$ and $(V', T + Y_{O'})$, and input these two samples separately into LVLMs, preserving the attention head activation values of the last token at each layer, denoted as A_v and $A_{v'}$ respectively. We define visual perception steering vector as the activation difference between visual input and distorted visual input according to Equation 2:

$$D_v = A_v - A_{v'} \tag{2}$$

Similarly, we apply PCA to D_v to reduce noise, thereby extracting the principal components most relevant to visual perception.

4.3 Dynamic Intervention for Inference

During the inference phase, for a given text input T and visual input V, we dynamically retrieve the most appropriate steering vector by computing semantic similarity between the input and each key in database as shown in Equation 3.

$$D_f = D_{\hat{i}}, \text{ where } \hat{i} = \arg\max_i \text{sim}(E(T), Key_i) \tag{3}$$

where $E(T)$ is the embedding representation of the input text, Key_i represents the key embedding for cluster i, and $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity function. This process identifies the most relevant truthfulness steering vector for the current input.

To achieve more precise control over model behavior, rather than intervening on all attention heads, we selectively target the most influential heads in both D_f and D_v.We define binary mask matrices M_f and M_v as Equation 4:

$$M_{\{f,v\}}^{(l,h)} = \begin{cases} 1, & \text{if } (l,h) \in \text{TopK}(\mathbf{D}_{\{f,v\}}, K) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where (l, h) denotes the h-th attention head in the l-th layer, $\mathbf{D}$ represents the sum of activation differences for each attention head in D and $\text{TopK}(\mathbf{D}_{\{f,v\}}, K)$ returns the indices of the K largest attention heads in either D_f or D_v, representing the most influential attention heads for truthfulness and visual perception respectively.

Building upon the standard attention mechanism, we modify the computation for layers where intervention is applied. Our intervention-enhanced computation is formulated as Equation 5:

$$\begin{aligned} \mathbf{x}^{(l+1)} = \mathbf{x}^{(l)} + \text{Concat}_{(0 \sim H)} \Big[& \text{Attn}^{(l,h)}(\mathbf{x}^{(l)}) \\ & + \alpha \cdot M_f^{(l,h)} \cdot D_f^{(l,h)} \\ & + \beta \cdot M_v^{(l,h)} \cdot D_v^{(l,h)} \Big] \cdot \mathbf{W}_o^{(l)} \end{aligned} \tag{5}$$

where $\mathbf{x}^{(l)}$ represents the hidden states at the l-th layer, H is the number of attention heads per layer, α and β are hyperparameters controlling the intervention strength for truthfulness and visual perception respectively. The binary masks ensure that interventions are only applied to the most influential attention heads, allowing for precise and targeted steering of the model's behavior.

5 Experimental Setup

5.1 Datasets and Evaluation Metrics

To comprehensively evaluate our proposed approach, we test our method on discriminative tasks, including MME Fu et al. (2023) and POPE Li et al. (2023c), as well as on open-ended generation tasks using CHAIR Rohrbach et al. (2018).

Model	Method	Existence↑	Count ↑	Position↑	Color↑	Total Scores↑
LLaVAv1.5	Regular	175.67	124.67	114.00	151.00	565.33
	VCD	184.66	138.33	128.67	153.00	604.66
	OPERA	180.67	133.33	123.33	155.00	592.33
	VAF	**195.00**	158.33	128.33	155.00	636.67
	AGLA	**195.00**	153.89	129.44	161.67	640.00
	ICT	190.00	**160.43**	128.67	170.00	649.10
	Ours	**195.00**	158.33	**133.33**	**173.33**	**659.99**
	Δ	↑19.33	↑33.66	↑19.33	↑22.33	↑94.66
QwenVL	Regular	155.00	127.67	131.67	173.00	587.33
	VCD	156.00	131.00	128.00	181.67	596.67
	VAF	165.00	**155.00**	**133.33**	175.00	628.33
	ICT	**180.00**	145.00	108.33	173.33	606.66
	Ours	170.00	145.00	**133.33**	**185.00**	**633.33**
	Δ	↑15.00	↑17.33	↑1.66	↑12.00	↑46.00

Table 1: Results on MME. The best results are shown in bold. Δ represents the improvement achieved by our method compared to the original model.

MME Fu et al. (2023) is a comprehensive evaluation benchmark for LVLMs, comprising 14 subtasks. For questions in this dataset, models are required to respond with either "yes" or "no". Following Yin et al. (2024) and Leng et al. (2024), we selected "existence," "count," "position," and "attribute" as the hallucination test sets. Consistent with Fu et al. (2023), we adopt the sum of accuracy and accuracy+ as the evaluation metrics.

POPE Li et al. (2023c) is a benchmark designed specifically to evaluate object hallucination. The benchmark features three sampling strategies of varying difficulty levels: random (randomly sampling nonexistent objects), popular (selecting frequently appearing objects), and adversarial (selecting objects that frequently co-occur with objects present in the image). We report Accuracy, Precision, Recall, F1 Score as the evaluation metrics.

CHAIR Rohrbach et al. (2018) is an open-ended generation task. This benchmark comprises 500 images sourced from MSCOCO Lin et al. (2014), where LVLMs are required to generate captions for the images, followed by evaluation of hallucinations present in these captions at sentence level CHAIR_S and image level CHAIR_I.

5.2 Baselines and Implementation Details

We validate the effectiveness of our method on mainstream LVLMs: LLaVAv1.5 7B Liu et al. (2024c) and QwenVL 7B Bai et al. (2023). We also compare our model with state-of-the-art methods: ICT Chen et al. (2025), AGLA An et al. (2025), VAF Yin et al. (2025), VTI Liu et al. (2025) , VCD Leng et al. (2024), and OPERA Huang et al. (2024).

Our method has three key parameters: α, β, and K. α and β respectively regulate the intensity of interventions for truthfulness and visual perception, while K refers to the intervention on the top K most active attention heads. We set the range of α and β to $\{0.5, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, and the range of K to $\{32, 64, 128, 256, 512, 1024\}$, and employ grid search to determine the parameters. All experiments are conducted on an NVIDIA RTX 4090(48GB) GPU.

6 Experiment

6.1 Results on MME

The results on the MME Fu et al. (2023) dataset are presented in Table 1. Our method demonstrates significant improvements of 94.66 and 46 points compared to the baseline models LLaVAv1.5 and QwenVL, respectively. On the LLaVAv1.5 model, our approach outperforms the existing state-of-

the-art method ICT Chen et al. (2025) by 10.89 points, while on QwenVL, it surpasses the current state-of-the-art method VAF Yin et al. (2025) by 5 points. Across all subtasks, we observe notable improvements over regular baselines, which can be attributed to our dynamic intervention mechanism that retrieves the most semantically similar steering vector for each query.

6.2 Results on POPE

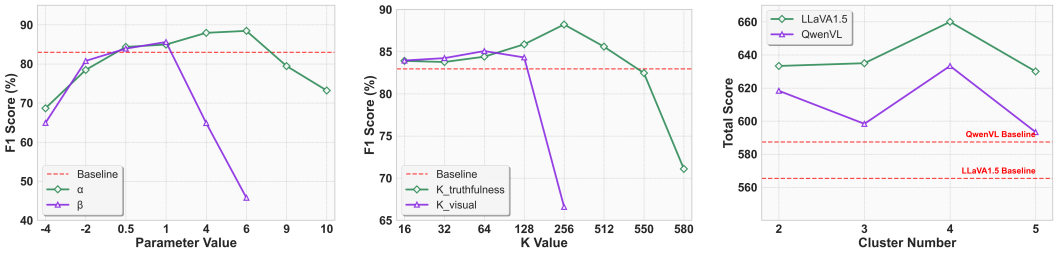
Dataset	Setting	Method	Accuracy ↑	Precision	Recall	F1 Score ↑
MSCOCO	LLaVAv1.5	Regular	81.38	88.04	72.78	79.65
		VCD	84.33	85.93	83.28	84.52
		OPERA	84.21	88.23	79.79	83.72
		VAF	86.90	89.43	83.77	86.47
		AGLA	85.82	93.78	76.83	84.44
		VTI	86.48	90.11	82.09	85.90
		ICT	**87.35**	-	-	**87.12**
		Ours	<u>86.81</u>(↑5.43)	87.23	86.57	<u>86.79</u>(↑10.14)
	QwenVL	Regular	83.71	93.30	72.69	81.70
		VCD	86.67	90.66	81.94	83.04
		OPERA	84.26	94.40	73.52	82.65
		AGLA	83.9	96.20	70.62	81.44
		VTI	85.18	91.31	78.18	84.08
		ICT	<u>87.53</u>	-	-	<u>86.98</u>
		Ours	**87.63**(↑3.92)	87.92	87.3	**87.65**(↑5.95)
GQA	LLaVAv1.5	Regular	78.33	79.33	79.13	79.13
		VCD	81.16	77.31	89.08	82.67
		OPERA	80.80	-	-	83.24
		VAF	83.67	81.50	88.00	84.50
		AGLA	<u>84.41</u>	84.63	84.67	84.55
		ICT	**85.27**	-	-	<u>85.50</u>
		Ours	**85.27**(↑6.94)	83.86	87.51	**85.63**(↑6.50)
	QwenVL	Regular	77.47	81.54	71.37	76.06
		VCD	82.48	81.73	83.93	82.77
		OPERA	82.74	-	-	82.68
		AGLA	81.14	86.87	73.53	79.63
		ICT	<u>83.28</u>	-	-	<u>83.26</u>
		Ours	**84.40**(↑6.93)	85.19	83.53	**84.32**(↑8.26)

Table 2: Results on POPE. Best results are in bold, and second-best values are underlined.

The experimental results of POPE Li et al. (2023c) are shown in Table 2. We conduct experiments on MSCOCO Lin et al. (2014) and GQA Hudson & Manning (2019) under random, popular, and adversarial settings. Table 2 presents the average results across these three settings, with detailed experimental results provided in the Appendix. Our method improved LLaVAv1.5's performance on MSCOCO by 5.43% in accuracy and 7.14% in F1 score, while for QwenVL, it achieved improvements of 3.92% in accuracy and 5.95% in F1 score. On GQA, our method enhances LLaVAv1.5 by 6.94% in accuracy and 6.5% in F1 score, and improves QwenVL by 6.93% in accuracy and 8.26% in F1 score. Compared to existing methods, our approach achieves best results in most cases, demonstrating its significant effectiveness in mitigating object hallucination. Notably, while the ICT Chen et al. (2025) method applies noise to objects in images to increase the LVLMs' attention to these objects, our method achieves superior performance in most cases without such specialized design elements.

6.3 Results on CHAIR

We evaluate our method on open-ended generation tasks, with experimental results on CHAIR Rohrbach et al. (2018) presented in Table 3. Our method reduces hallucinations by 20.2 at the



(a) Impact of α and β. (b) Impact of K. (c) Impact of Clusters.

Figure 3: Impact of key hyperparameters

sentence level (CHAIR_S) and by 3.8 at the image level (CHAIR_I). Compared to existing methods, our approach reduces sentence-level hallucinations by 5 points over the state-of-the-art method VTI Liu et al. (2025), and matching VTI's performance on image-level hallucinations. In summary, our method achieves significant improvements in hallucination mitigation on both discriminative tasks and open-ended generation tasks.

Method	$\text{CHAIR}_S\downarrow$	$\text{CHAIR}_I\downarrow$
Regular	51.0	15.2
VCD	51.0	14.9
OPERA	47.0	14.6
AGLA	43.0	14.1
VTI	35.8	**11.1**
Ours	**30.8**(↓**20.2**)	11.4(↓3.8)

Table 3: Results on CHAIR.

Method	CHAIR		POPE	
	$C_S\downarrow$	$C_I\downarrow$	Acc↑	F1↑
Ours	**30.8**	**11.4**	**81.70**	**82.47**
w/o visual vector	34.2	11.7	81.67	82.42
w/o truthfulness vector	42.4	13.2	81.40	82.01
w/o both	51.0	15.2	75.08	76.06

Table 4: Ablation studies on CHAIR and POPE. C_S represents CHAIR_S, C_I represents CHAIR_I.

6.4 Further Analysis

6.4.1 Ablation Studies

To demonstrate the effectiveness of using both truthfulness steering vector and visual perception steering vector, we compare the results when utilizing only one intervention at a time. We conduct experiments on LLaVAv1.5 using the CHAIR and POPE. As shown in Table 4, 'w/o visual vector' indicates intervention with only the truthfulness steering vector, while 'w/o truthfulness vector' indicates intervention with only the visual perception steering vector. We observe that even when using only one intervention method, there is a notable improvement compared to the Regular baseline (w/o both). Furthermore, each intervention method exhibits hallucination mitigation effects on both discriminative and generation tasks. The optimal results are achieved when the two interventions are combined.

6.4.2 Effect of Dynamic Intervention

To validate our designed strategy of dynamically invoking the truthfulness steering vector based on semantics, we compare our method with combining all truthfulness steering vectors into a single fixed steering vector for intervention. We conduct experiments on QwenVL and LLaVAv1.5 using the MME, with results shown in Figure 4. We observe that dynamically invoking steering vectors based on semantics achieves optimal performance across all subtasks. When using a fixed steering vector, the improvement is smaller than with our method, and it even underperforms the original model on the Position subtask of QwenVL, which demonstrates the necessity of our designed dynamic invocation strategy.

6.4.3 Impact of Hyperparameters

In this section, we investigate the impact of key parameters α, β, and K on the experimental results. Here, α and β control the intervention strength, while K denotes the number of attention heads receiving intervention. Experiments conducted on QwenVL using the POPE GQA random subset are shown in Figure 3. Figure 3a illustrates the relationship between F1 score and parameters α and β. When α and β are negative, we observe a decrease in F1 score, which effectively represents intervention in the opposite direction, pushing activations toward hallucination. As α and β increase, F1 score exhibits an upward trend; however, when α and β become excessively large, F1 score shows a precipitous decline, indicating that the model's fundamental capabilities become impaired. Figure 3b shows how F1 varies with the number of intervened attention heads, revealing similar patterns for both truthfulness and visual perception attention heads. Few intervened heads produce minimal impact with no significant F1 improvement. As intervention extends to more heads, F1 score increases, but excessive intervention causes a dramatic decline in F1, indicating degradation of model performance.

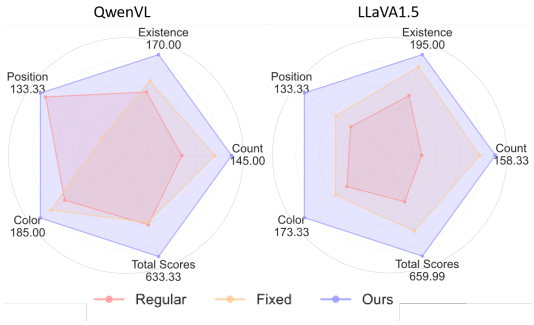


Figure 4: Effect of dynamic intervention.

6.4.4 Impact of Clusters

In this section, we investigate the impact of cluster quantity on the performance of our proposed method. We vary the number of clusters across $\{2, 3, 4, 5\}$ and conduct experiments on both QwenVL and LLaVAv1.5 using the MME benchmark. The experimental results are presented in Figure 3c. We observe that both LLaVAv1.5 and QwenVL achieve optimal performance when the number of clusters is set to 4. When the cluster count is insufficient, the semantic granularity becomes too coarse for effective representation. Conversely, when too many clusters are employed, the sample size within each cluster diminishes, leading to less stable steering vectors.

6.4.5 Analysis of Generality

To verify the generalizability of our method, we tested our approach on ScienceQA Lu et al. (2022) which is subject-based VQA dataset and ViQuAE Lerner et al. (2022) which is a knowledge-based VQA dataset. The accuracy are shown in Table 5. Our method also achieved significant improvements on these datasets. These datasets are completely different from the dataset types we used to construct the steering vector, which demonstrates the generalizability of our method.

Method	ScienceQA		ViQuAE	
	LLaVAv1.5	QwenVL	LLaVAv1.5	QwenVL
Regular	52.75	46.41	43.38	50.09
Ours	**62.27**	**48.04**	**56.00**	**54.08**

Table 5: Generality on ScienceQA and ViQuAE.

7 Conclusion

This paper proposes dynamic multimodal activation steering, a training-free plug-and-play approach to mitigate hallucination in LVLMs by dynamically intervening in attention head activations. The experimental results on multiple benchmarks demonstrate the effectiveness of our method, with LLaVA v1.5 achieving a remarkable 94.66-point improvement on MME and an average of 6.82-point increase on POPE, outperforming existing SOTA methods. We compare the experimental performance of our proposed semantic-dynamic strategy for steering vector selection against random selection and fixed steering vector approaches, demonstrating the effectiveness and necessity of semantic-dynamic steering vector selection.

REFERENCES

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29915–29926, 2025.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

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

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4209–4221, 2025.

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.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 958–979, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL https://arxiv.org/abs/2305.06500.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023. doi: 10.48550/ARXIV.2306.13394. URL https://doi.org/10.48550/arXiv.2306.13394.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, et al. Efficient multimodal large language models: A survey. *arXiv preprint arXiv:2405.10739*, 2024.

Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 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*, pp. 13872–13882, 2024.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. Viquae, a dataset for knowledge-based visual question answering about named entities. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pp. 3108–3120, 2022.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024a.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023b.

Xiaoqi Li, Mingxu Zhang, Yiran Geng, Haoran Geng, Yuxing Long, Yan Shen, Renrui Zhang, Jiaming Liu, and Hao Dong. Manipllm: Embodied multimodal large language model for object-centric robotic manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18061–18070, 2024b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne 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*, pp. 292–305, 2023c.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pp. 740–755. Springer, 2014.

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

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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*, pp. 26296–26306, 2024c.

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.

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.

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

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*, pp. 4035–4045, 2018.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. In *Proceedings of the 1st Workshop on Taming Large Language Models: Controllability in the era of Interactive Assistants!*, pp. 11–23, 2023.

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

Weixuan Wang, Jingyuan Yang, and Wei Peng. Semantics-adaptive activation intervention for llms via dynamic steering vectors. *arXiv preprint arXiv:2410.12299*, 2024a.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 15840–15853, 2024b.

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*, pp. 14625–14634, 2025.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

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

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

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness, 2024b. URL https://arxiv.org/abs/2405.17220.

Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia P Sycara, and Yaqi Xie. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jianfei Zhao, Feng Zhang, Xin Sun, and Chong Feng. Mitigate language priors in large vision-language models by cross-images contrastive decoding. *arXiv e-prints*, pp. arXiv–2505, 2025.

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*, 2024.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1624–1633, 2025.

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.

A APPENDIX

A.1 AI WRITING ASSISTANCE STATEMENT

Large language models were utilized solely for minor linguistic improvements, including enhanced phrasing and clarity. These tools played no role in content generation, experimental design, data analysis, or interpretation. The authors are entirely responsible for all ideas, results, and conclusions presented in this paper.

A.2 MORE DETAILS ON CHAIR

In this paper, we report CHAIR_S and CHAIR_I as evaluation metrics. The calculation of CHAIR_S and CHAIR_I is shown in Equation 6, where we set the maximum number of new tokens to 512 in our experiments.

$$\begin{aligned} \text{CHAIR}_S &= \frac{|\{\text{sentences with hallucinated objects}\}|}{|\{\text{all sentences}\}|} \\ \text{CHAIR}_I &= \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|} \end{aligned} \tag{6}$$

A.3 RESULTS ON POPE

The complete experimental results on POPE are presented in Table 6. Our method achieves significant improvements across all three experimental settings: random, popular, and adversarial.

Dataset	Setting	Method	Accuracy ↑	Precision	Recall	F1 Score ↑
MSCOCO	Random	Regular	83.29	92.13	72.80	81.33
		VCD	87.73	91.42	83.28	87.16
		Ours	**90.03**	90.51	90.03	**90.02**
	Popular	Regular	81.88	88.93	72.80	80.06
		VCD	85.38	86.92	83.28	85.06
		Ours	**87.33**	89.16	85.00	**87.03**
	Adversarial	Regular	78.96	83.06	72.75	77.57
		VCD	80.88	79.45	83.29	81.33
		Ours	**83.07**	82.04	84.67	**83.33**
GQA	Random	Regular	83.73	87.16	79.12	82.95
		VCD	86.65	84.85	89.24	86.99
		Ours	**89.57**	88.92	90.40	**89.60**
	Popular	Regular	78.17	77.64	79.12	78.37
		VCD	80.73	76.26	89.24	82.24
		Ours	**84.53**	83.51	86.07	**84.77**
	Adversarial	Regular	75.08	73.19	79.16	76.06
		VCD	76.09	70.83	88.75	78.78
		Ours	**81.70**	79.15	86.07	**82.47**

Table 6: Results on LLaVAv1.5. The best results are shown in bold.

A.4 RESULTS ON AMBER

We conduct an evaluation of LLaVA v1.5 on the AMBER Wang et al. (2023). AMBER contains both discriminative tasks and generative tasks. The experimental results are shown in Table 7. Our method outperforms existing methods on both discriminative and generative tasks, achieving significant effects in hallucination mitigation.

Method	Discriminative		Generative		
	Acc↑	F1↑	CHAIR↓	Hal↓	AMBER SCORE↑
Regular	67.4	71.2	11.6	47.7	79.80
VCD Leng et al. (2024)	68.1	71.1	9.8	43.8	80.65
ICD Wang et al. (2024b)	70.3	73.4	8.8	38.7	82.3
IBD Zhu et al. (2025)	69.2	72.2	9.8	42.2	81.2
DeFG Zhang et al. (2025)	70.2	73.0	9.1	39.9	81.95
CICD Zhao et al. (2025)	71.1	73.1	6.6	34.8	83.25
Ours	**81.9**	**87.2**	**4.9**	**20.9**	**90.01**

Table 7: Results on AMBER. The best results are shown in bold.

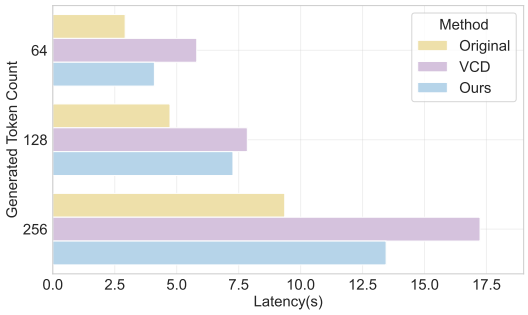


Figure 5: Effect of dynamic intervention.

A.5 Scalability Analysis Across Model Sizes

To verify that our method has hallucination mitigation effects for models of different sizes, we select the discriminative task dataset MME and the generative task dataset CHAIR on LLaVA1.5 7B and 13B models. The experimental results show in Table 8 that our model achieves significant effects for models of different sizes in both discriminative and generative tasks.

Model	Method	MME					CHAIR	
		Existence↑	Count ↑	Position↑	Color↑	Total Scores↑	$CHAIR_S$↓	$CHAIR_I$↓
LLaVAv1.5 7B	Regular	175.67	124.67	114.00	151.00	565.33	51.0	15.2
	Ours	**195.00**	**158.33**	**133.33**	**173.33**	**659.99**	**30.8**	**11.4**
LLaVA1.5 13B	Regular	**185.00**	131.67	95.00	175.00	586.67	45.0	11.8
	Ours	**185.00**	**158.33**	**103.33**	**180.00**	**626.66**	**38.0**	**10.8**

Table 8: Scalability Analysis Across Model Sizes on MME and CHAIR. The best results are shown in bold.

A.6 Analysis of Inference Speed

In this section, we investigate DMAS's inference speed. We use LLaVA1.5 7B and set the generation content lengths to {64, 128, 256} respectively, then compare the inference speed of our method with the original model and VCD. The experimental results show in Figure 5 that our method has faster inference speed compared to the decoding method VCD. VCD's inference latency is almost twice that of the original model, but our model achieves good hallucination mitigation effects while adding only a small amount of inference time.

A.7 Case Study

To intuitively demonstrate the hallucination mitigation effectiveness of our method, we conduct case studies on LLaVA v1.5. We utilize cases from MME and CHAIR datasets, with results shown

in Figure 6. Our method effectively mitigates multimodal hallucination issues across both VQA tasks and image captioning tasks. For VQA tasks, we present various question types, demonstrating our method's effectiveness in reducing hallucinations at different levels including object, attribute, relation, and count. For image captioning tasks, our method not only generates fewer hallucinations but also maintains the quality of the output content.

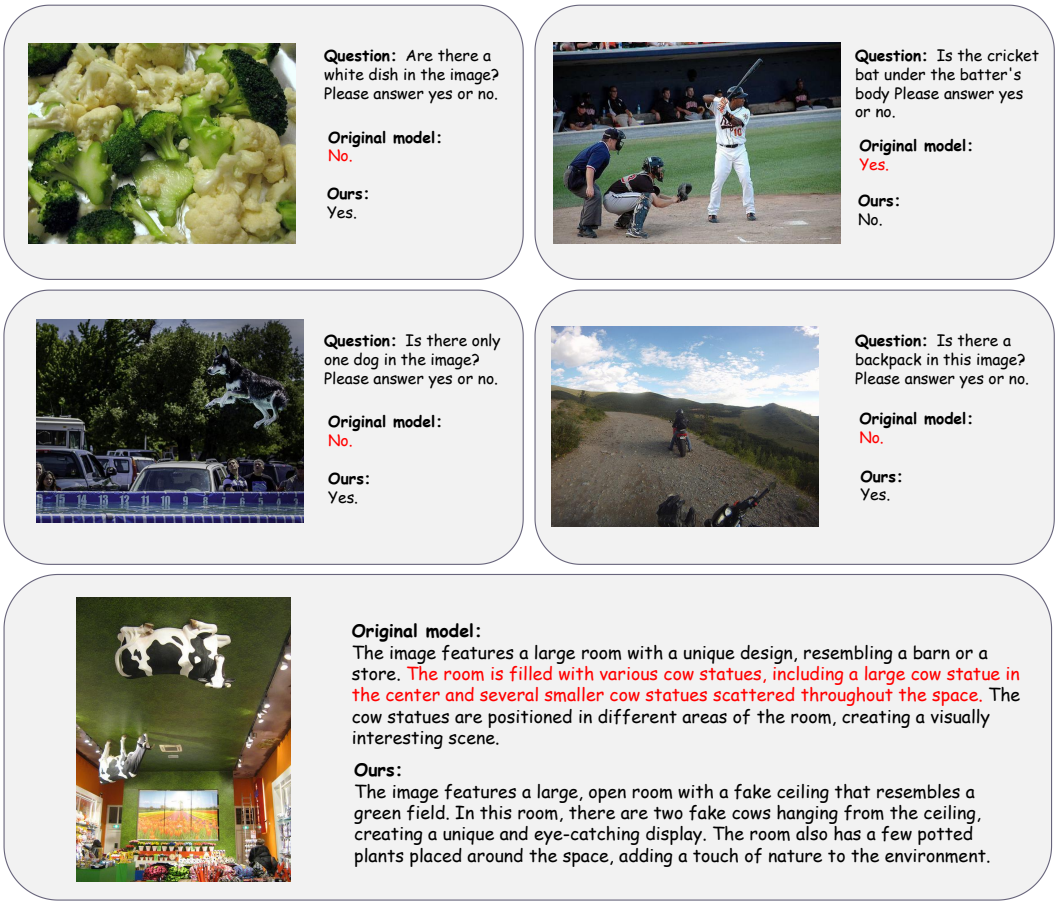


Figure 6: Case study on LLaVAv1.5.