ASCD: Attention-Steerable Contrastive Decoding for Reducing Hallucination in MLLM

Anonymous ACL submission

Abstract

Multimodal Large Language Model (MLLM) often suffer from hallucinations. They over-rely on partial cues and generate incorrect responses. Recently, methods like Visual Contrastive Decoding (VCD) and Instruction Contrastive Decoding (ICD) have been proposed to mitigate hallucinations by contrasting predictions from perturbed or negatively prefixed inputs against original outputs. In this work, we uncover that methods like VCD and ICD fundamentally influence the model's internal attention dynamics. This observation suggests that their effectiveness may not stem merely from surface-level modifications to logits but from deeper shifts in attention distribution. Inspired by this insight, we propose an *attention-steerable contrastive decoding* framework that directly intervenes in the model's attention mechanisms to offer a more principled approach to mitigating hallucinations. Specifically, we introduce *positive* and *negative steering* as two complementary directions for adapting the model's internal attention distributions. Rather than passively adjusting logits – as it is commonly done – our method dynamically modulates attention pathways within the contrastive decoding process. This enables selective enhancement or suppression of visual feature contributions in a structured manner. Furthermore, we propose a dynamic selection mechanism to identify *text-centric* heads – those that predominantly attend to text over visual features – for targeted positive steering, as well as a complementary mechanism to select the most critical visual tokens for negative steering, enabling more effective attention adjustments. Our experiments across multiple MLLM architectures (e.g., LLaVA-1.5 7B, LLaVA-NeXT 7B, Phi2-SigLIP) and diverse decoding methods (greedy search, beam search, nucleus sampling) demonstrate that our approach significantly reduces hallucinations and improves the performance on benchmarks such as POPE, CHAIR, and MMHAL-BENCH, while simultaneously enhancing performance on standard VQA benchmarks, including MMMU, MM-VET, SCIENCEQA, TEXTVQA, and GQA.

1 Introduction

Recent advances in large language models (LLMs) (Yang et al., 2024; Touvron et al., 2023; Abdin et al., 2024; Raffel et al., 2023; Brown et al., 2020; Devlin et al., 2019) have led to impressive results in a wide array of natural language processing tasks. Building on these successes, researchers have extended LLMs by visual inputs that enable multimodal large language models (MLLMs) such as LLaVA (Liu et al., 2023b, 2024b). These MLLMs can handle complex tasks like image captioning (Anderson et al., 2018), visual question answering (Agrawal et al., 2016), and multimodal dialogue (Das et al., 2017). Existing approaches (Dai et al., 2023; Liu et al., 2023b, 2024b; Zhou et al., 2024; Chen et al., 2023a; Alayrac et al., 2022) show remarkable potential to bridge the gap between vision and language.

Despite these achievements, MLLMs often inherit a critical limitation from LLMs: the tendency to produce *hallucinations* (Huang et al., 2024b; Bai et al., 2024; Liu et al., 2024a). These hallucinations arise when a model over-relies on partial or misleading cues, generating responses that are incorrect or do not correspond to the provided input.

To mitigate hallucinations, two general strategies have emerged: *training-phase* interventions and *inference-phase* interventions. In the training phase, auxiliary supervision (Chen et al., 2023b) or reinforcement learning (Ben-Kish et al., 2024) can help align model outputs with factual or human-preferred references. However, these approaches require additional data or complex reward modeling, which may be costly or infeasible in certain scenarios. In contrast, *inference-phase* methods (Zhou et al., 2024; Zhao et al., 2024; Deng et al., 2024; Wang et al., 2024; Leng et al., 2023) aim to

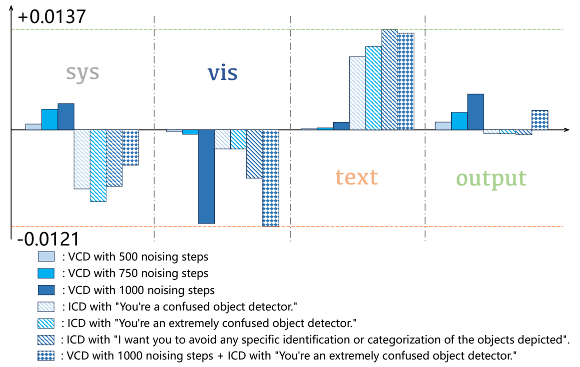

Figure 1: **Impact of VCD and ICD on attention distribution.** We conduct an image description task using LLaVA-1.5 on 500 randomly sampled images from the COCO dataset while monitoring the internal attention distribution within the LLM component. We compare the changes in attention under different settings of Visual Contrastive Decoding (VCD), Instruction Contrastive Decoding (ICD), and their combination, relative to the original LLaVA model. The x-axis represents different attention categories: system tokens (**sys**), visual tokens (**vis**), textual tokens (**text**), and output tokens (**output**). The y-axis indicates the attention difference relative to the original model. VCD (solid blue bars) reduces attention to visual tokens while slightly increasing attention to textual tokens, with a stronger effect as the number of noising steps increases. ICD (hatched bars) exhibits a similar trend, further decreasing visual attention and increasing text attention, where stronger negative prefixes (see text in the legend) result in a more pronounced shift. When combining VCD and ICD (dotted bars), the reduction in visual attention is further amplified, while the focus on textual tokens increases. These findings indicate that the effectiveness of VCD and ICD originates from underlying shifts in the model's attention distribution rather than solely from the contrastive decoding process.

correct or filter erroneous outputs without retraining. *Contrastive decoding* is particularly appealing as it leverages negatively perturbed or prefixed inputs to steer the model away from hallucinations in a training-free manner. Two notable recent methods for contrastive decoding are Visual Contrastive Decoding (VCD) (Leng et al., 2023) that perturbs an input image (*e.g.*, via noising) to generate a "negative result" of logits, which is then subtracted from the original logits to suppress hallucinations, and second, Instruction Contrastive Decoding (ICD) (Wang et al., 2024) that prepends a negative prefix to the prompt (*e.g.*, "You are a confused object detector") to generate a signal that shifts the model's predictions away from hallucinated content. Both methods offer a lightweight, yet effective approach to reducing hallucinations. However, upon closer examination, we find that these methods construct contrasting branches through surface-level modifications – either perturbing the image (VCD) or prefixing the prompt (ICD) – without explicitly addressing the underlying cause of hallucinations. *Attention steering* like OPERA and PAI (Liu et al., 2024c; Huang et al., 2024a) is also a common inference-phase remedy to reduce hallucination. However, PAI introduces the notion of "text inertia" – the tendency of an MLLM to keep generating text-driven content even when the image is removed – but does not articulate why steering the attention matrix is the necessary lever to overcome this inertia.

In our experiments (Fig. 1), we observe that both VCD and ICD consistently cause fundamental shifts in the internal attention distribution: they tend to reduce attention on visual tokens and amplify attention on textual tokens. This insight raises a natural question: *why not directly steer the attention mechanism itself?* To this end, we propose an *Attention-Steerable Contrastive Decoding (ASCD)* framework to manipulate attention. Specifically, the attention modification is integrated into a contrastive decoding pipeline to either enhance visual cues or to suppress negative signals. We further develop a dynamic head-selection mechanism to identify "text-centric" heads that disproportionately focus on textual cues, enabling more targeted positive adjustments. In parallel, we introduce a complementary mechanism that restricts negative steering to only the most critical visual tokens, ensuring that suppression is applied solely where necessary to mitigate hallucinations while preserving essential visual details. In summary, our contributions are as follows: (1) We analyze how recent contrastive decoding methods (VCD, ICD) create "negative samples" that fundamentally alter attention; (2) We propose an *attention-steerable contrastive decoding* method that explicitly modulates attention distributions to offer a more principled way to mitigate hallucinations in the inference phase; (3) We faithfully reproduce VCD and ICD to ensure fair comparison with prior work. Across three representative MLLM backbones (LLaVA-1.5 7B, LLaVA-NeXT 7B, and Phi2-SigLIP), three decoding schemes (greedy, nucleus, and beam search), and three hallucination-focused benchmarks (Rohrbach et al., 2019; Li et al., 2023b; Sun et al., 2023) (POPE, CHAIR, MMHAL-BENCH), our approach consistently reduces hallucinations and strengthens visual grounding. At the same time, it improves performance on standard VQA benchmarks (Yue et al., 2024; Yu et al., 2024; Lu et al., 2022; Singh et al., 2019; Hudson and Manning, 2019), including MMMU, MM-VET, ScienceQA, TextVQA, and GQA whereas other methods suffer from degraded performance on these benchmarks.

2 Related Work

Multimodal Large Language Models. Multimodal Large Language Models (MLLMs) have significantly advanced the field of artificial intelligence by integrating vision and language understanding, enabling a wide range of vision-language tasks (Dai et al., 2023; Zhu et al., 2023; Liu et al., 2024b, 2023b; Alayrac et al., 2022; Chen et al., 2023a; Zhou et al., 2024). These models typically follow a two-stage training paradigm: (1) large-scale pretraining on web-scale image-text pairs (Liu et al., 2023b; Li et al., 2023a) to learn cross-modal representations, and (2) visual instruction tuning (Liu et al., 2023a; Bi et al., 2025) on task-specific datasets to enhance multimodal instruction-following capabilities. While this paradigm has led to substantial improvements in vision-language reasoning, MLLMs still face key challenges, such as hallucination – where the model generates content that is inconsistent with the given visual input. (Huang et al., 2024b; Bai et al., 2024; Liu et al., 2024a).

Mitigating Hallucinations in MLLMs. Hallucination in MLLMs is particularly pronounced in open-ended generation tasks, where models may produce content that is not aligned with the provided visual input (Huang et al., 2024a; Jing et al., 2024; Zhang et al., 2023). Some approaches focus on the mitigation of data bias, scaling-up of vision resolution, and alignment optimization. Lovenia et al. (2024) introduce a technique that mines 95,000 negative samples by replacing original categories, attributes, or quantity information with similar but incorrect alternatives. This fine-grained approach effectively enriches the contrastive signal during training, thereby enhancing the model's robustness. Chen et al. (2024) propose InternVL, which scales the vision encoder up to 6 billion parameters and processes images with widths ranging from 1,664 to 6,144 pixels. While this method improves visual detail and alignment, it requires significant computational resources for pretraining with large-scale data. Sun et al. (2023) employ Reinforcement Learning from Human Feedback (RLHF) (Stiennon et al., 2022) to align different modalities during training. This optimization strategy leads to a reduction in hallucinations by better integrating visual and textual cues.

Contrastive Decoding Approaches. Recent work has explored contrastive decoding as an effective, training-free means to mitigate hallucinations. For instance, Leng et al. (2023) introduced Visual Contrastive Decoding (VCD), which perturbs the input image to generate a negative logit branch that is subtracted from the original predictions, while Wang et al. (2024) employs a negative prompt to steer outputs away from hallucinated content. Huo et al. (2024) leverages a Context and Text-aware Token Selection (CT2S) strategy to selectively retain the most informative vision tokens in early decoder layers, thereby amplifying beneficial multimodal context and suppressing spurious hallucinations.

3 Preliminaries

Modern MLLMs integrate text and visual inputs based on powerful encoders that enable merging the modalities into a unified representation that is processed by a multi-layer Transformer. While these models enable producing coherent responses, they heavily rely on internal attention mechanisms that dictate how visual and textual cues are combined. As discussed in Section 3.2, subtle variations in these attention distributions can significantly impact the generated output. This observation motivates our approach: by explicitly modulating attention, we aim to enhance visual grounding and mitigate hallucinations.

3.1 MLLM Formulation

We consider a multimodal large language model (MLLM) that processes an image $\mathbf{I}$ and a text prompt $\mathbf{x} = \{x_1, \ldots, x_N\}$ to generate an output sequence $\mathbf{y} = \{y_1, \ldots, y_M\}$ in an autoregressive manner. Let θ denote the model parameters. Formally, the model maximizes:

$$\mathbf{y}^* = \arg\max_{\mathbf{y}} \prod_{t=1}^{M} p_\theta\Big(y_t \,\Big|\, \mathbf{I}, \mathbf{x}, y_{<t}\Big), \tag{1}$$

where $y_{<t}$ denotes all previously generated tokens.

Embeddings. A unified input is obtained from encoded image and embedded text:

$$\mathbf{Z} = [f_v(\mathbf{I}); f_t(\mathbf{x})]. \tag{2}$$

Transformer Architecture. The MLLM processes $\mathbf{Z}$ through L Transformer layers (Vaswani

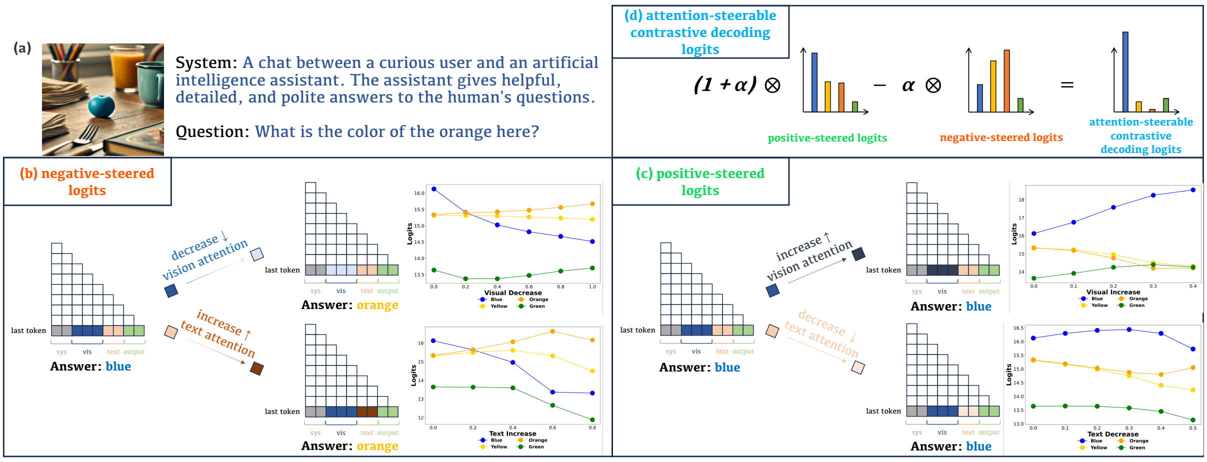

Figure 2: **A motivating example of proactive attention steering in a visually ambiguous scenario.** **(a)** shows the conversation context where the "orange" is actually tinted blue. **(b)** shows how the logits vary based on negative-steering. **(c)** shows how the logits vary based on positive-steering. **(d)** illustrates how attention-steerable contrastive decoding, which combines both negative and positive steering in a unified framework, reduce hallucinations and produce perception-driven answers.

et al., 2023):

$$\mathbf{H}^{(l)} = \text{TransformerLayer}^{(l)}\big(\mathbf{H}^{(l-1)}\big), \quad \mathbf{H}^{(0)} = \mathbf{Z}. \tag{3}$$

Output Prediction. The final hidden state $\mathbf{h}_t^{(L)}$ is mapped to a probability distribution over the vocabulary:

$$p_\theta(y_t \mid \mathbf{I}, \mathbf{x}, y_{<t}) = \text{Softmax}\Big(\mathbf{h}_t^{(L)} W^P\Big), \tag{4}$$

where W^P is the output projection matrix.

3.2 Proactive Steering of Attention

In Figure 1, we show how Visual Contrastive Decoding (VCD) and Instruction Contrastive Decoding (ICD) indirectly alter attention distributions. Building on this insight, we now ask: *what if we explicitly steer the model's attention?* Figure 2 provides a motivating example, illustrating how actively modulating attention can influence the final logits distribution.

Consider a simple query: *"What is the color of the orange here?"* The conversation context (Figure 2a) is based on LLaVA-1.5 7B, with a provided image in which the "orange" fruit appears to be tinted blue. We experiment with two distinct attention-steering scenarios: *negative-steered logits* (Figure 2b) and *positive-steered logits* (Figure 2c). In each case, we proportionally adjust the visual or textual attention before finalizing the output distribution.

In the *negative-steered* branch, we reduce attention to visual tokens or boost attention to the textual tokens. As shown in the histogram of logits, the model reduces its reliance on the visual input, causing it to fall back more heavily on the LLM's inherent priors. As a result, it is more likely to generate answers that align with typical linguistic associations rather than the actual content of the image – insisting the color is "orange". Conversely, the *positive-steered* branch increases attention to visual tokens or downgrades textual tokens, making the model more sensitive to the actual (albeit unexpected) color in the image. This leads the model to answer "blue" with higher probability.

In addition to these unidirectional adjustments, we further integrate *attention steering* into the contrastive decoding framework. Instead of using the original logits directly (as in VCD or ICD), we inject the attention-modulated logits. Mathematically, we redefine the contrastive decoding formulation by replacing the original logits adjustment with a positively steered version:

$$p_\theta^{\text{final}} = (1+\alpha)p_\theta^{\text{pos-steered}} - \alpha p_\theta^{\text{neg-steered}}, \tag{5}$$

where $p_\theta^{\text{pos-steered}}$ and $p_\theta^{\text{neg-steered}}$ represent the output logits modified by positively or negatively steered attention.

By integrating contrastive decoding with explicit attention manipulation, our attention-steerable contrastive decoding framework (Figure 2d), sharpens the output distribution which enhances the likelihood of the correct response, while reducing the impact of competing distractors.

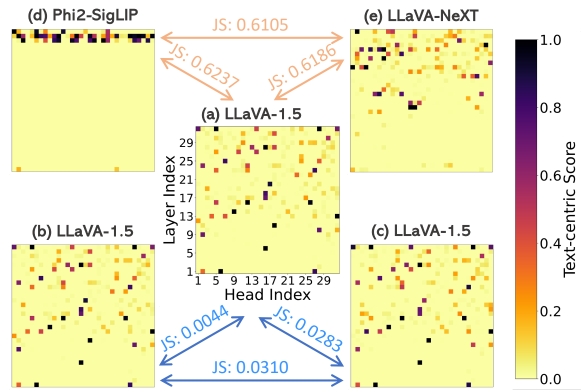

Figure 3: **Distribution of text-centric heads across different models and experiment settings.** Each heatmap visualizes how frequently a given head occurs among the most text-focused heads. The panel in the center (a) show the result of LLaVA-1.5 with a generation length of 64 tokens; (b) and (c) show results of the same model with longer generation (512 tokens) and a different image set. Despite these changes, LLaVA-1.5 exhibits minimal JS divergence, which indicates consistent text-centric heads. In contrast, Phi2-SigLIP (d) and LLaVA-NeXT (e) deviate substantially from LLaVA-1.5, revealing model-specific attention biases and higher JS divergence.

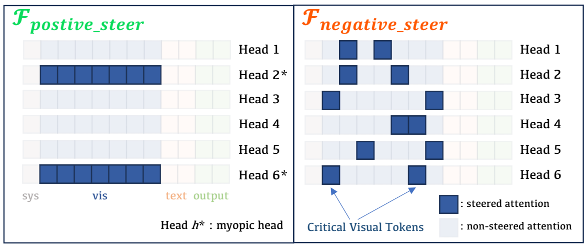

Figure 4: **Illustration of positive and negative steering.** Left: text-centric heads are boosted (*positive_steer*) to emphasize visual content; Right: a small set of critical visual tokens is suppressed (*negative_steer*), inducing a stronger contrastive effect. These selective adjustments work in tandem to reduce hallucinations and improve grounding.

3.3 Text-centric Heads

Previously, we have highlighted the impact of adjusting attention. In this section, we discuss *which* heads in the model are most prone to over-reliance on textual cues. To this end, we conduct an experiment to identify "text-centric" heads, i.e., those with disproportionately high text-to-visual attention ratios, and examine their consistency under different generation conditions and image sets. The experiment setup is detailed in Appendix A.

Results and Observations. Figure 3 shows the resulting heatmaps F for multiple models and generation settings. The panel in the center (a) corresponds to LLaVA-1.5 on $N = 500$ images with a generation length of 64 tokens. The two heatmaps at the bottom show results of the same model but with either an increased generation length to 512 tokens (b, bottom left), or using a different set of 500 images (c, bottom right). Despite these changes, the distribution of top text-focus heads remains visually similar, and the small Jensen–Shannon (JS) divergences confirm that these text-centric heads are largely invariant under different sampling conditions for *the same model*.

In contrast, the Phi2-SigLIP (d, top-left) and LLaVA-NeXT (e, top-right) panels deviate significantly from LLaVA-1.5 even under the same experiment settings, with higher JS divergence. This suggests that each model has its own unique set of heads that consistently favor textual attention over visual cues. However, *within* a single model, the text-centric heads persist across varied prompts, image sets, and generation lengths.

Implications. The consistent presence of the text-centric heads within the same model indicates that certain heads are inherently prone to focusing on textual signals rather than visual content. In Section 4.2 we describe how this insight can be leveraged to selectively target the problematic heads when applying our *positive steering* strategy. Rather than uniformly amplifying attention across all heads, we concentrate on those that are most responsible for text-dominant attention, thereby avoiding unnecessary modifications to heads that are well-balanced in their visual-textual focus.

4 Methodology

In this section, we present our *attention-steerable contrastive decoding* framework, which explicitly modulates the model's attention to mitigate hallucinations. Our approach has two stages: (1) *Text-centric Head Selection*, which identifies the heads most prone to text-centric bias, and (2) *Attention Steering*, where we apply positive steering to text-centric heads and negative steering to a small subset of visually critical tokens. We then integrate these adjusted logits for generation into a contrastive decoding pipeline.

4.1 Text-centric Head Selection

As detailed in Algorithm 1, we start by identifying the most *text-centric* heads using a small reference dataset (e.g., 500 images) for a task (e.g., image description). For each sample, we compute the

Algorithm 1 Text-centric Head Selection (Offline)

Require: Reference dataset $\{\mathbf{I}_1, \ldots, \mathbf{I}_N\}$, MLLM with L layers and H heads per layer, Final text-centric head count κ_{TCH}
Ensure: $\mathcal{H}_{\text{POS}}$ (set of selected text-centric heads)

Initialize a global counter $F \in R^{L \times H}$ to zeros
for $i \leftarrow 1$ to N **do**
  Run MLLM on image $\mathbf{I}_i$ (e.g., image description)
  for all head (r, c) in layer-head grid **do**
    Compute attention ratio:

$$Q_i(r, c) = \frac{\text{textAttn}(r, c)}{\text{visAttn}(r, c)}$$

  end for
  Identify top-32 indices of Q_i (largest ratios) and store in $\mathcal{I}_i$
  for all $(r, c) \in \mathcal{I}_i$ **do**
    $F(r, c) \leftarrow F(r, c) + 1$
  end for
end for
Sort all heads (r, c) in descending order by $F(r, c)$
Select top κ_{TCH} heads:

$$\mathcal{H}_{\text{POS}} \leftarrow \text{first } \kappa_{\text{TCH}} \text{ heads in sorted list}$$

return $\mathcal{H}_{\text{POS}}$

Algorithm 2 Attention-Steerable Contrastive Decoding

Require: Image $\mathbf{I}$, Text-centric heads $\mathcal{H}_{\text{POS}}$ (from Algorithm 1), Critical vis-token count κ_{VIS}, Steer strengths $\alpha_{\text{POS}}, \alpha_{\text{NEG}}$, Contrastive weight α, Truncation threshold β, MLLM with L layers and H heads per layer
Ensure: p_θ^{final} (final logits from ASCD)

Step 1: Forward Pass with Positive Steering
for $l \leftarrow 1$ to L **do**
  for $h \leftarrow 1$ to H **do**
    Compute attention matrix $\mathbf{A}_h^{(l)}$
    if $(l, h) \in \mathcal{H}_{\text{POS}}$ **then**
      $\mathbf{A}_h^{(l)} \leftarrow \mathbf{A}_h^{(l)} + \alpha_{\text{POS}} \, |\mathbf{A}_h^{(l)}|$
    end if
  end for
  Normalize $\mathbf{A}^{(l)}$ and continue
end for
Obtain logits $p_\theta^{\text{pos-steered}}$

Step 2: Forward Pass with Negative Steering
for $l \leftarrow 1$ to L **do**
  for $h \leftarrow 1$ to H **do**
    Compute attention matrix $\mathbf{A}_h^{(l)}$
    Identify top-κ_{VIS} critical visual tokens $\mathcal{V}$
    for all $v \in \mathcal{V}$ **do**
      $\mathbf{A}_h^{(l)}(v) \leftarrow \mathbf{A}_h^{(l)}(v) - \alpha_{\text{NEG}} \, |\mathbf{A}_h^{(l)}(v)|$
    end for
  end for
  Normalize $\mathbf{A}^{(l)}$ and continue
end for
Obtain logits $p_\theta^{\text{neg-steered}}$

Step 3: Contrastive Decoding with Truncation
$p_\theta^{\text{raw}} \leftarrow (1 + \alpha) \, p_\theta^{\text{pos-steered}} - \alpha \, p_\theta^{\text{neg-steered}}$
$\text{cutoff} \leftarrow \log(\beta) + \max(p_\theta^{\text{raw}})$
$p_\theta^{\text{final}} \leftarrow p_\theta^{\text{raw}}.\text{masked\_fill}(p_\theta^{\text{pos-steered}} < \text{cutoff}, -\infty)$
return p_θ^{final}

ratio of textual attention to visual attention (Eq. 6 in Appendix A) and take the top 32 heads with the highest ratio. We accumulate these counts over all samples, then choose the top κ_{TCH} heads as "text-centric". This step is motivated by our finding (Section 3.3) that certain heads consistently favor textual content over visual cues.

4.2 Attention Steering

Text-centric Head Awareness and Critical Visual Token Selection. As shown in Figure 4, we refine our method by incorporating text-centric head selection for positive steering and critical token identification for negative steering. Specifically, given the selected text-centric heads, we *positively steer* them by increasing their attention weights with a strength of α_{POS}. Figure 5a highlights how targeted steering in text-centric heads improves the positive steering effectiveness. Simultaneously, we apply *negative steering* to the top κ_{VIS} most critical visual tokens – *those receiving the highest aggregate attention across heads* – reducing their attention scores by α_{NEG} across *all* heads. Through this strategy, we deliberately obscure only the most pivotal cues – this targeted suppression is sufficient to induce a strong hallucination effect in the negative branch, leading to improved contrastive decoding compared to a blanket suppression of all visual tokens. In Figure 5b, we demonstrate the impact of selectively applying negative steering to critical visual tokens.

Integration with Contrastive Decoding with Truncation. Building on the attention-steering process, we first obtain two output distributions: $p_\theta^{\text{pos-steered}}$ from the positively steered branch and $p_\theta^{\text{neg-steered}}$ from the negatively steered branch. We then combine these into contrastive decoding with a truncation mechanism, as detailed in the Step 3 of Algorithm 2. This process not only reinforces visually grounded predictions but also effectively mitigates the influence of spurious textual biases.

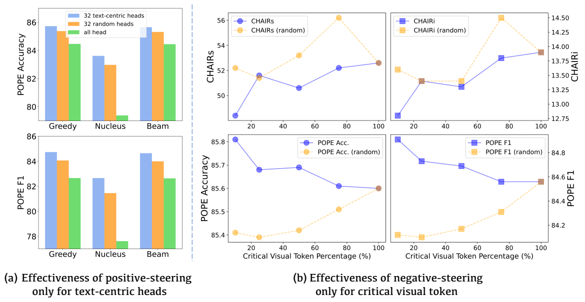

Figure 5: **Comparative effectiveness of selective attention steering. (a):** Positive steering applied *only* to text-centric heads (32 heads with the highest text-to-visual ratio) outperforms random or blanket head selection across various decoding strategies (Greedy, Nucleus, Beam), leading to higher POPE Accuracy and F1. **(b):** Negative steering focused on a small subset of critical visual tokens, integrated with contrastive decoding, significantly reduces CHAIR metrics (less hallucination) and boosts POPE metrics compared to randomly suppressing visual tokens of the same number. These results validate that *targeted* attention modulation on text-centric heads (for positive steering) and critical visual tokens (for negative steering) yields stronger hallucination mitigation and more grounded responses.

Model	Decoding	Method	CHAIRs (↓)	CHAIRi (↓)
LLaVA-1.5 7B	greedy	Orig	53.2	13.5
		VCD	56.8	15.2
		ICD	52.8	13.2
		ASCD	**35.6**	**8.6**
	nucleus	Orig	59.0	17.4
		VCD	59.8	16.6
		ICD	57.4	15.6
		ASCD	**43.6**	**11.3**
	beam	Orig	54.8	15.3
		VCD	58.8	16.4
		ICD	52.6	13.9
		ASCD	**40.8**	**10.1**
LLaVA-NeXT 7B	greedy	Orig	31.6	7.5
		VCD	37.2	9.7
		ICD	32.8	8.4
		ASCD	**21.8**	**7.0**
	nucleus	Orig	30.4	8.0
		VCD	40.4	10.4
		ICD	39.4	9.9
		ASCD	**21.2**	**6.7**
	beam	Orig	34.0	8.5
		VCD	36.6	9.1
		ICD	31.8	7.6
		ASCD	**21.0**	**6.5**
Phi2-SigLIP	greedy	Orig	29.0	6.9
		VCD	39.4	9.6
		ICD	33.4	7.7
		ASCD	**21.8**	**5.4**
	nucleus	Orig	36.0	9.8
		VCD	36.0	8.1
		ICD	37.0	9.4
		ASCD	**26.0**	**8.0**
	beam	Orig	30.4	6.9
		VCD	36.0	8.4
		ICD	31.0	7.0
		ASCD	**24.6**	**5.7**

Table 1: **CHAIR Evaluation Results.** Lower CHAIRs and CHAIRi values indicate better performance in reducing hallucination. The best values for each metric within a model-decoding combination are highlighted in **bold**.

5 Experiments

To evaluate the effectiveness of our attention-steerable contrastive decoding framework in mitigating hallucinations in MLLMs, we conduct a range of experiments. This includes three diverse benchmarks – **CHAIR**, **POPE**, and **MMHal-Bench** – each designed to assess different aspects of object hallucinations. To ensure the broad applicability and robustness of our approach, we also test it on three representative models – **LLaVA-1.5 7B**, **LLaVA-NeXT 7B**, **Phi2-SigLIP**, and employ three different decoding strategies: **greedy search**, **nucleus sampling**, and **beam search**. Details of the experimental settings are provided in Appendix B. Furthermore, we evaluate performance on standard VQA benchmarks including **MMMU**, **MM-VET**, **ScienceQA**, **TextVQA**, and **GQA** to verify that the proposed method preserves – rather than diminishes – the model's original visual understanding.

It is important to note that current benchmarks for evaluating multimodal models are highly variable. For example, baseline models such as LLaVA-1.5 7B often report different metric values between different papers. Moreover, the CHAIR metric relies on random image sampling, which further complicates direct comparisons between papers. To address these issues, we faithfully *reproduced* both VCD and ICD using the parameters specified in their original papers and repositories, ensuring that our evaluations are conducted under consistent conditions. This allows for a more reliable comparison between our method and existing approaches.

CHAIR. Table 1 shows the CHAIR metrics (CHAIRs and CHAIRi), which measure object hallucination in image captioning. Across all models and decoding strategies, ASCD consistently achieves *lower* CHAIR values than Orig, VCD, or ICD, which illustrates ASCD's effectiveness at mitigating object-level hallucinations.

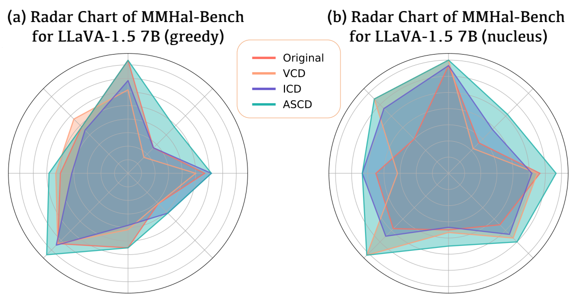

Figure 6: **Radar charts of MMHal-Bench results.** Each axis represents a different evaluation dimension in MMHal-Bench, and a larger enclosed area indicates better overall performance.

Model	Decoding	Method	Popular Acc	Popular F1	Random Acc	Random F1	Adv. Acc	Adv. F1	Avg. Acc	Avg. F1
LLaVA-1.5 7B	greedy	Orig	85.83	84.35	86.67	85.51	**83.60**	82.31	85.37	84.06
		VCD	84.67	83.59	86.43	85.53	81.70	80.93	84.27	83.35
		ICD	83.47	80.85	83.68	81.51	82.07	79.56	83.07	80.64
		PAI	-	-	-	-	-	-	85.82	85.79
		ASCD	**87.20**	**86.69**	**89.48**	**89.09**	82.90	**82.97**	**86.53**	**86.25**
	nucleus	Orig	83.47	81.74	84.78	83.44	80.83	79.53	83.03	81.57
		VCD	83.83	82.61	85.12	84.12	80.97	80.18	83.31	82.30
		ICD	82.63	79.85	82.68	80.43	81.07	78.58	82.13	79.62
		PAI	-	-	-	-	-	-	81.72	82.87
		ASCD	**86.47**	**85.56**	**87.94**	**87.26**	**82.83**	**82.39**	**85.75**	**85.07**
	beam	Orig	85.87	84.39	86.70	85.55	**83.63**	82.36	85.40	84.10
		VCD	84.43	83.30	86.19	85.31	82.20	81.30	84.27	83.30
		ICD	83.47	80.83	83.61	81.43	82.03	79.51	83.04	80.59
		PAI	-	-	-	-	-	-	86.33	85.89
		ASCD	**87.20**	**86.69**	**89.48**	**89.09**	82.87	**82.95**	**86.52**	**86.24**
LLaVA-NeXT 7B	greedy	Orig	83.97	81.77	85.09	83.26	82.73	80.64	83.93	81.89
		VCD	84.87	83.12	86.19	84.84	**83.53**	**81.89**	**84.86**	83.28
		ICD	84.53	82.63	85.70	84.15	83.10	81.33	84.44	82.70
		ASCD	**84.90**	**83.30**	**86.39**	**85.09**	83.27	81.82	84.85	**83.40**
	nucleus	Orig	81.73	79.26	83.61	81.75	79.87	77.81	81.74	79.61
		VCD	84.20	82.51	84.78	83.26	81.67	80.07	83.55	81.95
		ICD	83.60	81.68	85.29	83.78	82.13	80.47	83.67	81.98
		ASCD	**84.60**	**82.86**	**86.19**	**84.77**	**83.27**	**81.65**	**84.69**	**83.09**
	beam	Orig	84.17	82.04	85.26	83.49	82.90	80.88	84.11	82.14
		VCD	84.67	82.86	86.19	84.81	83.13	81.42	84.66	83.03
		ICD	84.57	82.68	85.74	84.19	83.13	81.37	84.48	82.75
		ASCD	**84.97**	**83.39**	**86.43**	**85.14**	**83.33**	**81.91**	**84.91**	**83.48**
Phi2-SigLIP	greedy	Orig	87.10	85.95	88.45	87.57	86.03	84.97	87.19	86.16
		VCD	86.00	85.14	87.97	87.37	84.70	84.09	86.22	85.53
		ICD	85.50	84.14	87.25	86.15	84.73	83.44	85.83	84.58
		ASCD	**87.77**	**86.74**	**88.90**	**88.14**	**86.77**	**85.81**	**87.81**	**86.90**
	nucleus	Orig	85.73	84.49	86.87	85.87	83.93	82.96	85.51	84.44
		VCD	85.60	84.72	86.91	86.21	84.30	83.64	85.60	84.86
		ICD	84.90	83.46	85.98	84.86	83.00	81.73	84.63	83.35
		ASCD	**87.50**	**86.41**	**88.52**	**87.69**	**86.33**	**85.29**	**87.45**	**86.46**
	beam	Orig	87.10	85.95	88.45	87.57	86.03	84.97	87.19	86.16
		VCD	86.43	85.63	87.90	87.36	84.57	83.92	86.30	85.64
		ICD	85.50	84.14	87.25	86.15	84.73	83.44	85.83	84.58
		ASCD	**87.77**	**86.74**	**88.90**	**88.14**	**86.77**	**85.81**	**87.81**	**86.90**

Table 2: **POPE Evaluation Results.** The best values for each metric within a model-decoding combination are highlighted in **bold**. If our ASCD achieves the second-best result, it is additionally marked with an underline.

POPE. Table 2 reports the accuracy and F1 scores under the POPE evaluation, which probes object presence with random, popular, and adversarial queries. Higher values indicate fewer hallucinations. Again, ASCD achieves the best or near-best performance in all cases. These gains persist across different model architectures, suggesting that attention steering is robust to model size and design variations.

MMHal-Bench. Figure 7 illustrates the radar charts of MMHal-Bench results for LLaVA-1.5 7B under greedy and nucleus decoding. Each axis represents a sub-dimension of the benchmark, and a larger area signifies better overall performance. ASCD exhibits the largest enclosed area, outperforming baseline, VCD, and ICD in most dimensions. This improvement aligns with the CHAIR and POPE findings, underscoring the benefit of selectively steering attention to reduce hallucinations.

Standard VQA Benchmarks. To verify that ASCD does not sacrifice a model's general visual-question-answering ability, it's evaluated on five widely-used VQA datasets. Across all three backbones and all decoding strategies, ASCD either matches or surpasses the original model on every dataset, while VCD and ICD consistently degrade performance as shown in the Appendix C.

Summary. Our experiments confirm that ASCD effectively reduces hallucinations and improves alignment with visual content, regardless of the model or decoding strategy employed.

6 Conclusion

We have shown that existing contrastive methods (e.g. VCD and ICD) inadvertently *shift* the internal attention distribution in MLLM, prompting us to investigate a more direct and principled way to modulate attention. We proposed an *attention-steerable contrastive decoding* framework that *positively steers* text-centric heads while *negatively steering* only the most critical visual tokens.

Our method consistently reduces hallucinations on **CHAIR**, **POPE**, and **MMHal-Bench**, outperforming both baseline and previous contrastive approaches with improved and uncompromised general VQA capability. By targeting precisely those heads and tokens, we effectively mitigate spurious textual biases while preserving essential visual context.

Limitations

A key limitation of ASCD is its incompatibility with FlashAttention. Because our method must dynamically modify the attention matrix at inference time, it cannot make use of the fused kernels, leading to higher memory consumption and slower decoding. A promising workaround is to distill the steering signal into the model during training: we can add an auxiliary loss – e.g., a KL-divergence term – that drives the native attention distribution to approximate the ASCD target distribution. If successful, the model would internalise the hallucination-mitigation behaviour, removing the need for on-the-fly edits and restoring FlashAttention speed-ups. We regard training-time attention regularization as a promising direction: it could distill the hallucination-robust behaviour discovered by training-free, attention-modified methods into the model itself, so that at inference the model retains this robustness while fully benefiting from FlashAttention's speed and memory efficiency.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, and 110 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2016. Vqa: Visual question answering. *Preprint*, arXiv:1505.00468.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. *Preprint*, arXiv:1707.07998.

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

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. 2024. Mitigating open-vocabulary caption hallucinations. *Preprint*, arXiv:2312.03631.

Jinhe Bi, Yifan Wang, Danqi Yan, Xun Xiao, Artur Hecker, Volker Tresp, and Yunpu Ma. 2025. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *Preprint*, arXiv:2502.12119.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, and 10 others. 2023a. Pali: A jointly-scaled multilingual language-image model. *Preprint*, arXiv:2209.06794.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Preprint*, arXiv:2312.14238.

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

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. *Preprint*, arXiv:1611.08669.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024a. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *Preprint*, arXiv:2311.17911.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. 2024b. Visual hallucinations of multi-modal large language models. *Preprint*, arXiv:2402.14683.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Preprint*, arXiv:1902.09506.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2024. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *Preprint*, arXiv:2408.02032.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. *Preprint*, arXiv:2311.01477.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *Preprint*, arXiv:2311.16922.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *Preprint*, arXiv:2305.10355.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

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

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

Shi Liu, Kecheng Zheng, and Wei Chen. 2024c. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *Preprint*, arXiv:2407.21771.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2024. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *Preprint*, arXiv:2310.05338.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Preprint*, arXiv:2209.09513.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. Object hallucination in image captioning. *Preprint*, arXiv:1809.02156.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. *Preprint*, arXiv:1904.08920.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. *Preprint*, arXiv:2009.01325.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *Preprint*, arXiv:2403.18715.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. *Preprint*, arXiv:2308.02490.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023. How language model hallucinations can snowball. *Preprint*, arXiv:2305.13534.

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

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024. Tinyllava: A framework of small-scale large multimodal models. *Preprint*, arXiv:2402.14289.

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

A Text-centric Heads Experiment Settings

The following setup applies to Section 3.3.

We select $N = 500$ images $\{\mathbf{I}_1, \dots, \mathbf{I}_N\}$ (from COCO) and run an MLLM (LLaVA-1.5) in an image description task. During each generation, we track the ratio of textual attention to visual attention for every head:

$$Q_i \in R^{R\times C}, \quad Q_i(r,c) = \frac{\text{textAttn}(r,c)}{\text{visAttn}(r,c)}, \tag{6}$$

where r and c index each head (for instance, $R = C = 32$). We then identify the top-k heads with the highest ratio values and mark them in a binary mask:

$$M_i(r,c) = \begin{cases} 1, & \text{if } (r,c) \in \text{top-}k \text{ indices of } Q_i, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Finally, we aggregate these masks across all N images:

$$F = \sum_{i=1}^{N} M_i, \tag{8}$$

so that $F(r,c)$ records how frequently head (r,c) appears among the most text-focused heads.

B Evaluation Settings

B.1 Baseline Models and Decoding Methods

We evaluate our proposed approach using three representative models and multiple decoding strategies to demonstrate its broad applicability and robustness.

LLaVA-1.5 7B is a minimalist yet efficient model that has served as the foundation for extensive studies in large multimodal modeling. **LLaVA-NeXT 7B** builds on LLaVA-1.5 with improvements in visual reasoning, higher input resolution, and enhanced world knowledge, resulting in superior performance on several benchmarks. **Phi2-SigLIP** leverages the Phi-2 backbone and a SigLIP-based vision tower, and is trained on the ShareGPT4V dataset, offering a compact alternative with competitive capabilities.

To assess the reliability and generalizability, we experiment with three decoding strategies: **greedy search**, **nucleus sampling**, and **beam search**. Greedy decoding yields deterministic outputs, while nucleus sampling and beam search enable for more diverse generation.

B.2 Datasets

We evaluate our approach on three hallucination-targeted benchmark datasets designed to probe object hallucination in multimodal large language models.

CHAIR. The Caption Hallucination Assessment with Image Relevance (CHAIR) metric quantifies the degree of hallucination in generated captions by measuring the fraction of objects mentioned that do not actually appear in the image. It is computed at both the instance-level (CHAIRi) and the sentence-level (CHAIRs), offering insight into how well a caption adheres to veridical image content.

POPE. The Polling-based Object Probing Evaluation (POPE) assesses hallucination by querying the model with binary questions (*e.g.*, "Is there a car in the image?"). By balancing queries about present and absent objects, and using different sampling strategies (random, popular, adversarial), POPE effectively reveals the influence of language priors on model predictions. This method provides a robust measure of object hallucination across multiple datasets such as MSCOCO, A-OKVQA, and GQA.

MMHal-Bench. MMHal-Bench is a new evaluation benchmark specifically designed to challenge large multimodal models in hallucination.

Comprising 96 difficult questions based on images, along with detailed ground-truth answers and image content annotations, MMHal-Bench offers a comprehensive testbed to assess the model's ability to produce veridical responses in ambiguous visual scenarios.

We also evaluate our approach on five standard VQA benchmarks to verify that our method does not compromise, and may even enhance, the original VQA capabilities.

MM-VET. The MM-VET benchmark evaluates large multimodal models on integrated vision–language abilities that go beyond conventional VQA. It defines six core capabilities—recognition, OCR, knowledge, language generation, spatial awareness, and math—and constructs 16 task types that require their various combinations. MM-Vet comprises 200 images paired with 218 open-ended questions; answers are graded by a GPT-4-based evaluator that yields unified scores, enabling fine-grained, per-capability diagnostics.

ScienceQA. The ScienceQA benchmark targets multimodal, curriculum-level science reasoning. It contains multiple-choice questions drawn from elementary- to high-school materials, each optionally paired with text passages, natural or diagrammatic images, or both. In this study, answers are graded for accuracy.

GQA. The GQA benchmark targets real-world compositional visual reasoning. Questions span objects, attributes, relations, spatial logic, comparisons, and multi-hop inference, making GQA a stringent yardstick for models that aspire to deep vision-language understanding rather than pattern-matching.

TextVQA. The TextVQA benchmark focuses on visual question answering that explicitly requires reading and reasoning over scene text. It contains 45,336 open-ended questions paired with 28,408 everyday images drawn from the Open Images dataset. Questions span practical tasks such as reading product labels, traffic signs, receipts, and clocks, so models must combine OCR, spatial grounding, and commonsense reasoning.

MMMU. The Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) benchmark gauges college-level, expert-style reasoning across six broad disciplines—Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering.

B.3 Hyperparameters

The experiments are conducted with the following hyperparameter settings to obtain the best result:

$\kappa_{\text{VIS}} = 0.1$, $\alpha_{\text{NEG}} = 1.0$

LLaVA-1.5: $\kappa_{\text{TCH}} = 32$, $\alpha_{\text{POS}} = 0.6$, $\alpha = 1.0$, and $\beta = 0.5$.

LLaVA-NeXT: $\kappa_{\text{TCH}} = 32$, $\alpha_{\text{POS}} = 0.7$, $\alpha = 1.0$, and $\beta = 0.7$.

Phi2-SigLIP: $\kappa_{\text{TCH}} = 128$, $\alpha_{\text{POS}} = 0.8$, $\alpha = 0.5$, and $\beta = 0.1$.

C Standard VQA Capability

Our approach substantially curbs hallucinations through explicit attention steering while simultaneously enhancing performance on standard VQA tasks, a trade-off that competing methods typically fail to avoid.

C.1 MM-VET

Method	MM-VET Score
Orig	31.2
VCD	30.3
ICD	33.2
ASCD	33.2

Table 3: MM-VET scores for different methods.

C.2 ScienceQA

Method	SQA Score
Orig	67.55
VCD	67.55
ICD	67.32
ASCD	69.51

Table 4: SQA scores for different methods.

C.3 GQA

Method	GQA Score
Orig	61.28
VCD	59.38
ICD	59.99
ASCD	61.27

Table 5: GQA scores for different methods.

C.4 TextVQA

Method	TextVQA Score
Orig	57.82
VCD	55.07
ICD	57.66
ASCD	57.91

Table 6: TextVQA scores for different methods.

C.5 MMMU

Method	MMMU Score
Orig	0.342
VCD	0.333
ICD	0.360
ASCD	0.348

Table 7: MMMU scores for different methods.

D Visualization Example

Instruction: Describe the image in details.

LLaVA-1.5 (Orig):
The image depicts a group of people walking along a sidewalk near a river. There are at least six people visible in the scene, with some of them walking closer to the water and others further away. A woman is pushing a stroller, and a man is walking with a child in a stroller. The sidewalk is lined with parked cars, with a total of nine cars visible in the scene. Some cars are parked closer to the water, while others are situated further away. The overall atmosphere of the scene is that of a pleasant day spent outdoors, with people enjoying the company of others and the beautiful surroundings.

LLaVA-1.5 (VCD):
The image features a beach scene with a man sitting on a four-wheeler, which is a type of all-terrain vehicle. The man is positioned near the water, possibly observing the waves or preparing to ride the vehicle on the beach. There are several other people in the scene, some of them closer to the water and others further away. A surfboard can be seen lying on the sand, indicating that surfing might be a popular activity in this area. Additionally, there is a boat visible in the background, adding to the beach atmosphere.

LLaVA-1.5 (ICD):
The image features a man riding a four-wheeler, also known as an ATV, on a sandy beach near the ocean. He is wearing a helmet for safety while enjoying his ride. The beach is lively, with several people scattered around, some closer to the water and others further away. In the background, there is a pier extending out into the ocean, providing a picturesque view for the beachgoers. The scene captures a fun and relaxing atmosphere, with people enjoying their time by the water.

LLaVA-1.5 (ASCD):
The image features a man riding a four-wheeler, which is a type of all-terrain vehicle, on a sandy beach near the ocean. The man is sitting on a four-wheeler, which is an off-road vehicle with large tires, and he is driving it on the beach. There are several other people in the background, some of whom are standing on the pier. The pier extends out into the ocean, providing a scenic view of the water. The beach is a popular spot for people to enjoy the outdoors and engage in various activities.

Figure 7: Oringinal LLaVA-1.5 drifts completely, describing a riverside sidewalk, parked cars, and a stroller that do not exist in the image. VCD and ICD correctly identify the beach and ATV but still hallucinate extra objects such as a surfboard and a boat (highlighted in yellow). Our ASCD caption mentions only entities that are actually present and omits non-existent items, illustrating the improved visual grounding achieved through attention-steerable contrastive decoding.