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

Yujun Wang[1], **Aniri**[2,3], **Jinhe Bi**[3,4], **Soren Pirk**[1], **Yunpu Ma**[3,4]
[1]Christian-Albrechts-Universität zu Kiel (CAU)
[2]Technical University of Munich
[3]Ludwig Maximilian University of Munich
[4]Munich Center for Machine Learning
yujun_wang_cn@hotmail.com, sp@informatik.uni-kiel.de, cognitive.yunpu@gmail.com

Abstract

Multimodal large language models (MLLMs) frequently hallucinate by over-committing to spurious visual cues. Prior remedies—Visual and Instruction Contrastive Decoding (VCD, ICD)—mitigate this issue, yet the mechanism remains opaque. We first empirically show that their improvements systematically coincide with *redistributions of cross-modal attention*. Building on this insight, we propose **Attention-Steerable Contrastive Decoding (ASCD)**, which *directly steers the attention scores during decoding*. ASCD combines (i) *positive steering*, which amplifies automatically mined *text-centric heads*—stable within a model and robust across domains—with (ii) *negative steering*, which dampens on-the-fly identified critical visual tokens. The method incurs negligible runtime/memory overhead and requires no additional training. Across five MLLM backbones and three decoding schemes, ASCD reduces hallucination on POPE, CHAIR, and MMHAL-BENCH by up to 38.2% while *improving* accuracy on standard VQA benchmarks, including MMMU, MM-VET, SCIENCEQA, TEXTVQA, and GQA. These results position attention steering as a simple, model-agnostic, and principled route to safer, more faithful multimodal generation.

Code — https://github.com/BroJunn/ASCD

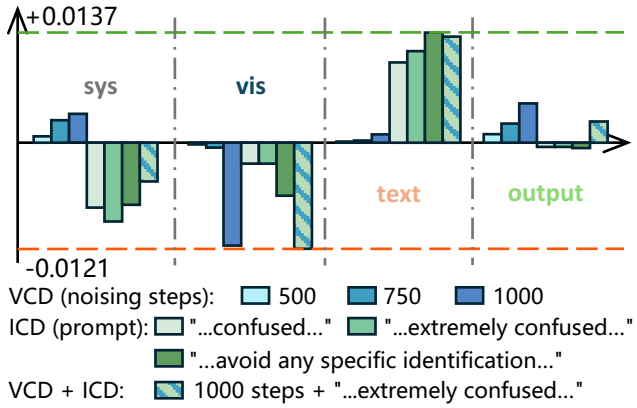

Figure 1: Impact of VCD and ICD on attention distribution. On 500 COCO images, we measure how Visual (VCD) and Instruction (ICD) Contrastive Decoding redistribute attention in LLaVA-1.5. Both techniques—and their combination—lower attention on visual tokens (*vis*) while raising it on textual tokens (*text*), with stronger perturbations yielding larger shifts.

Introduction

Recent advances in large language models (LLMs) (Team 2024b; Touvron et al. 2023; Team 2024a; 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 with visual inputs that enable multimodal large language models (MLLMs) such as LLaVA (Liu et al. 2023b, 2024a). These MLLMs can handle complex tasks like image captioning (Anderson et al. 2018), visual question answering (Agrawal et al. 2016; Yu et al. 2025a; Chen et al. 2025b; Yu et al. 2025b), and multimodal dialogue (Das et al. 2017). Existing approaches (Dai et al. 2023; Liu et al. 2023b, 2024a; Zhou et al. 2024; Chen et al. 2023a; Alayrac et al. 2022; Bi et al. 2024) 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. 2024b). 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, Chen, and Hooi 2024; Wang et al. 2024; Leng et al. 2023) aim to 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 man-

ner. 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 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, IBD and PAI (Liu, Zheng, and Chen 2024; Zhu et al. 2024; Huang et al. 2024a) is also a common inference-phase remedy to reduce hallucinations. 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.

To motivate our approach, we first quantify how VCD and ICD reshape a model's internal attention. As evidenced by Fig. 1, both techniques produce a systematic reallocation of attention from visual tokens to 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 both enhance visual cues and 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 five representative MLLM backbones (LLaVA-1.5 7B, LLaVA-NeXT 7B, Phi2-SigLIP, LLaVA-1.5 13B and Qwen2.5-VL-Instruct), 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.

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. 2024a, 2023b; Bai et al. 2025; Zhou et al. 2024; Rong et al. 2025; Chen et al. 2025a; Lu et al. 2024a; Lu, Liu, and Kong 2023; Lu et al. 2024b, 2025). 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 (Huang et al. 2024b; Bai et al. 2024; Liu et al. 2024b; Huang et al. 2024a).

Mitigating Hallucinations in MLLMs. 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. Bi et al. (2024) propose a representation steering method that effectively mitigates hallucination in multimodal models.

Contrastive Decoding Approaches. Recent work has explored contrastive decoding as an effective, training-free means to mitigate hallucinations (Xiao et al. 2025). 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) leverage a Context and Text-aware Token Selection strategy to selectively retain the most informative vision tokens in early decoder layers, thereby amplifying beneficial context and suppressing spurious hallucinations.

Preliminaries

Modern MLLMs integrate text and visual inputs based on powerful encoders that enable the merging of the modalities into a unified representation that is processed by a multi-layer Transformer.

While these models enable the production of coherent responses, they heavily rely on internal attention mechanisms that dictate how visual and textual cues are combined. As discussed in the previous section, subtle variations in these attention distributions can significantly impact the generated output. This observation motivates our approach: by explic-

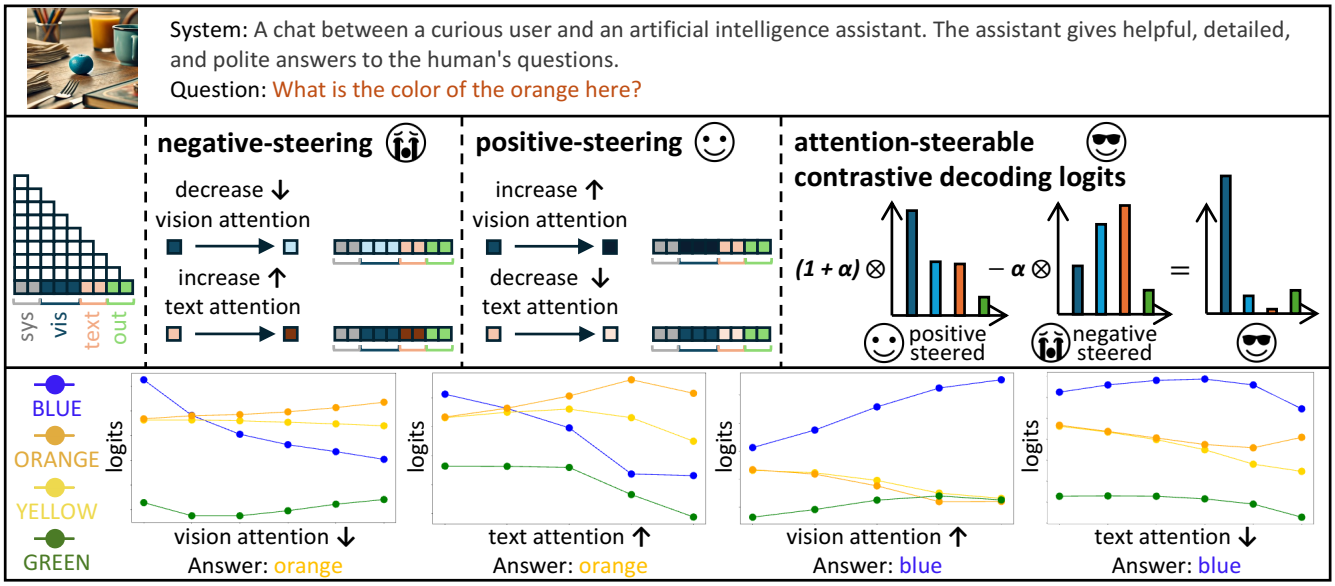

Figure 2: A motivating example of proactive attention steering in a visually ambiguous scenario. Top: Conversation context in which the "orange" appears blue-tinted. Middle: Effects of *negative steering* (decrease vision attention / increase text attention) and *positive steering* (increase vision attention / decrease text attention); ASCD contrasts the two steered logits to suppress hallucination and produce the perception-consistent answer. Bottom: Color-token logits change with the steering strength for visual and textual attention, corresponding to the steering above.

itly modulating attention, we aim to enhance visual grounding and mitigate hallucinations.

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:

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

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

Transformer Backbone. The input is processed by L Transformer blocks, and each block contains H attention heads. We denote the unnormalized attention score matrix of the head h in layer l by $\mathbf{A}_h^{(l)}$.

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 logit distribution.

Consider a simple query: *"What is the color of the orange here?"* The conversation context (Figure 2) 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* and *positive-steered logits*. 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 that 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 for the positive branch 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:

$$p_\theta^{\text{final}} = (1+\alpha)p_\theta^{\text{pos-steered}} - \alpha p_\theta^{\text{neg-steered}}, \tag{2}$$

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 2 right) sharpens the output distribution,

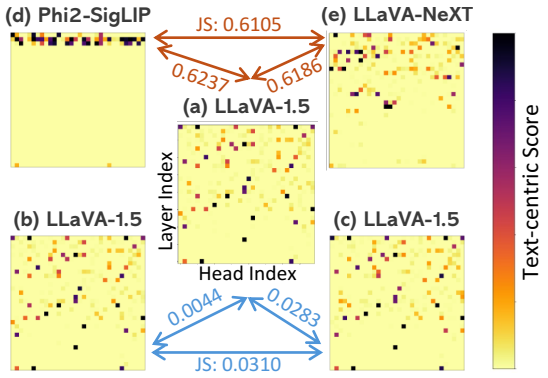

Figure 3: The stability of text-centric head distribution. Each heatmap visualizes how frequently a given head occurs among the most text-focused heads. LLaVA-1.5(a) remains stable across *generation length*(b) and *image set*(c), whereas Phi2-SigLIP(d) and LLaVA-NeXT(e) shift markedly.

enhances the likelihood of the correct response, while reducing the impact of competing distractors.

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.

Text-centric Heads are Model-specific

Having established the impact of attention adjustments, we now discuss *which* heads in the model are most prone to over-reliance on textual cues. Here, 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 experimental setup is detailed in the supplementary material (Appendix, Text-Centric Heads Experiment Settings).

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-focused 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

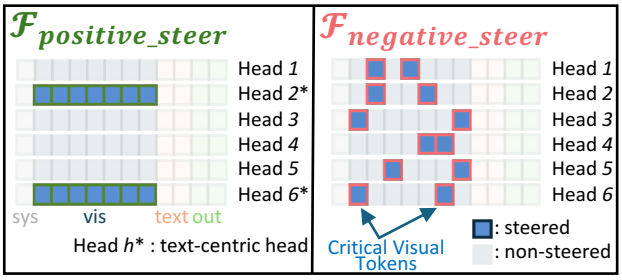

Figure 4: Illustration of positive and negative steering. Left: text-centric heads are boosted (*positive_steer*) to emphasize visual content; Right: a small set of critical visual tokens is suppressed (*negative_steer*), inducing a stronger contrastive effect. These selective adjustments work in tandem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

LLaVA-1.5 even under the same experimental 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 the next subsection, we describe how this insight can be leveraged to selectively target the problematic heads when applying our *positive steering* strategy.

Robustness across Data Domains. To further assess the robustness of *model-specific* text-centric heads, we repeat the profiling on extremely out-of-domain X-ray data. Details are provided in the supplementary material ("Extended Analysis of Text-centric Heads on Medical Data Domain"). The resulting heatmaps and Jensen–Shannon divergences show that, within each model, the *same subset of heads* remains text-centric despite the shift from COCO photographs to chest X-rays, indicating strong domain robustness.

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 ratio of textual attention to visual attention 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 previous finding that certain heads consistently favor textual content over visual cues.

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

Algorithm 1: Text-Centric Head Selection (Offline)

Input: Reference image set $\{\mathbf{I}_1, \ldots, \mathbf{I}_N\}$; MLLM with L layers and H heads per layer; desired text-centric head count κ_{TCH}

Output: $\mathcal{H}_{\text{POS}}$

Initialize Global Statistics:

Initialize counter tensor $F \leftarrow \mathbf{0}^{L \times H}$

Vote Accumulation Over Reference Set:

foreach $\mathbf{I}_i \in \{\mathbf{I}_1, \ldots, \mathbf{I}_N\}$ **do**

  Run MLLM on $\mathbf{I}_i$ to obtain cross-modal attentions

  foreach $(r, c) \in \{1, \ldots, L\} \times \{1, \ldots, H\}$ **do**

    `// Loop over all heads`

    $Q_i(r, c) \leftarrow \dfrac{\text{textAttn}(r, c)}{\text{visAttn}(r, c)}$

  $\mathcal{I}_i \leftarrow$ indices of the **top-32** Q_i values

  foreach $(r, c) \in \mathcal{I}_i$ **do**

    $F(r, c) \leftarrow F(r, c) + 1$

Head Selection:

Sort heads (r, c) by $F(r, c)$ in descending order

$\mathcal{H}_{\text{POS}} \leftarrow$ top κ_{TCH} heads

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

a strength of α_{POS}. Figure 5a highlights how targeted steering in text-centric heads improves the positive steering effectiveness. Simultaneously, we perform *negative steering* on the κ_{VIS} visual tokens that draw the most attention—those with the highest head-averaged score. For the current query position, we define

$$s(v) = \frac{1}{H} \sum_{h=1}^{H} A_h^{(L)}(v), \qquad v \in \{1, \ldots, V\}. \tag{3}$$

The κ_{VIS} tokens with the largest $s(v)$ values form the critical set $\mathcal{V}_{\text{crit}}$; every attention entry to any $v \in \mathcal{V}_{\text{crit}}$ is then downscaled in *all* heads by $\alpha_{\text{NEG}} |A_h^{(L)}(v)|$.

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. We first compute two output distributions: p_θ^{pos} from the positively steered branch and p_θ^{neg} from the negatively steered branch. Step 3 of Algorithm 2 then fuses them through a contrastive-decoding rule with truncation, producing the final logits. This fusion amplifies visually grounded evidence while suppressing spurious text-only cues, thereby reducing hallucinations.

Algorithm 2: Attention-Steerable Contrastive Decoding (ASCD)

Input: Image $\mathbf{I}$; text-centric heads $\mathcal{H}_{\text{POS}}$; critical visual-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

Output: p_θ^{final}

1. Positive Steering Pass:

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**

      $\mathbf{A}_h^{(l)} \leftarrow \mathbf{A}_h^{(l)} + \alpha_{\text{POS}} |\mathbf{A}_h^{(l)}|$

  Normalize $\mathbf{A}^{(l)}$ and continue

Obtain logits p_θ^{pos}

2. Negative Steering Pass:

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 $\mathcal{V}_{\text{crit}}$

    foreach $v \in \mathcal{V}_{crit}$ **do**

      $\mathbf{A}_h^{(l)}(v) \leftarrow \mathbf{A}_h^{(l)}(v) - \alpha_{\text{NEG}} |\mathbf{A}_h^{(l)}(v)|$

  Normalize $\mathbf{A}^{(l)}$ and continue

Obtain logits p_θ^{neg}

3. Contrastive Decoding and Truncation:

$p_\theta^{\text{raw}} \leftarrow (1 + \alpha)\, p_\theta^{\text{pos}} - \alpha\, p_\theta^{\text{neg}}$

cutoff $\leftarrow \log(\beta) + \max(p_\theta^{\text{raw}})$

$p_\theta^{\text{final}} \leftarrow p_\theta^{\text{raw}}.\text{masked\_fill}(p_\theta^{\text{pos}} < \text{cutoff}, -\infty)$

return p_θ^{final}

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 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 the supplementary material (Appendix, Evaluation Settings). 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. In addition, an *extended* evaluation on the larger **LLaVA-1.5 13B** and the modern **Qwen2.5-VL-Instruct** is conducted to verify scalability and generality on a subset of benchmarks.

It is important to note that current benchmarks for evaluating MLLMs are highly variable. For example, baseline models such as LLaVA-1.5 often report different metric values between different papers. Moreover, CHAIR relies on random sampling, which further complicates direct compar-

Model	Decoding	Method	CHAIRs (↓)	CHAIRi (↓)	POPE-Acc (↑)	POPE-F1 (↑)
LLaVA-1.5 7B	greedy	Orig	53.2	13.5	85.37	84.06
		VCD	56.8	15.2	84.27	83.35
		ICD	52.8	13.2	83.07	80.64
		PAI	-	-	85.82	85.79
		ASCD	**35.6 (33.1%)**	**8.6 (36.3%)**	**86.53**	**86.25**
	nucleus	Orig	59.0	17.4	83.03	81.57
		VCD	59.8	16.6	83.31	82.30
		ICD	57.4	15.6	82.13	79.62
		PAI	-	-	81.72	82.87
		ASCD	**43.6 (26.1%)**	**11.3 (35.1%)**	**85.75**	**85.07**
	beam	Orig	54.8	15.3	85.40	84.10
		VCD	58.8	16.4	84.27	83.30
		ICD	52.6	13.9	83.04	80.59
		PAI	-	-	86.33	85.89
		ASCD	**40.8 (25.5%)**	**10.1 (34.0%)**	**86.52**	**86.24**
LLaVA-NeXT 7B	greedy	Orig	31.6	7.5	83.93	81.89
		VCD	37.2	9.7	**84.86**	83.28
		ICD	32.8	8.4	84.44	82.70
		ASCD	**21.8 (31.0%)**	**7.0 (6.7%)**	<u>84.85</u>	**83.40**
	nucleus	Orig	30.4	8.0	81.74	79.61
		VCD	40.4	10.4	83.55	81.95
		ICD	39.4	9.9	83.67	81.98
		ASCD	**21.2 (30.3%)**	**6.7 (16.3%)**	**84.69**	**83.09**
	beam	Orig	34.0	8.5	84.11	82.14
		VCD	36.6	9.1	84.66	83.03
		ICD	31.8	7.6	84.48	82.75
		ASCD	**21.0 (38.2%)**	**6.5 (23.5%)**	**84.91**	**83.48**
Phi2-SigLIP	greedy	Orig	29.0	6.9	87.19	86.16
		VCD	39.4	9.6	86.22	85.53
		ICD	33.4	7.7	85.83	84.58
		ASCD	**21.8 (24.8%)**	**5.4 (21.7%)**	**87.81**	**86.90**
	nucleus	Orig	36.0	9.8	85.51	84.44
		VCD	36.0	8.1	85.60	84.86
		ICD	37.0	9.4	84.63	83.35
		ASCD	**26.0 (27.8%)**	**8.0 (18.4%)**	**87.45**	**86.46**
	beam	Orig	30.4	6.9	87.19	86.16
		VCD	36.0	8.4	86.30	85.64
		ICD	31.0	7.0	85.83	84.58
		ASCD	**24.6 (19.1%)**	**5.7 (17.4%)**	**87.81**	**86.90**
LLaVA-1.5 13B	greedy	Orig	51.2	12.6	85.52	84.12
		ASCD	**33.0(35.5%)**	**8.0(36.5%)**	**87.78**	**87.40**
	nucleus	Orig	51.4	14.5	83.95	82.54
		ASCD	**35.7(30.5%)**	**9.5(34.5%)**	**87.19**	**86.46**
Qwen-2.5-VL-Instruct	greedy	Orig	31.2	7.5	87.72	86.64
		ASCD	**24.4(21.8%)**	**6.4(14.7%)**	**88.91**	**88.27**
	nucleus	Orig	34.8	8.6	87.34	86.14
		ASCD	**26.9(22.7%)**	**6.3(26.7%)**	**88.45**	**87.68**

Table 1: CHAIR and POPE Evaluation Results. Lower CHAIRs and CHAIRi values indicate better performance in reducing hallucinations. POPE performance is reported as the mean accuracy and F1 score. The best values for each metric within a model-decoding combination are highlighted in bold. If ASCD ranks second, the best is bold while the ASCD score is <u>underlined</u>.

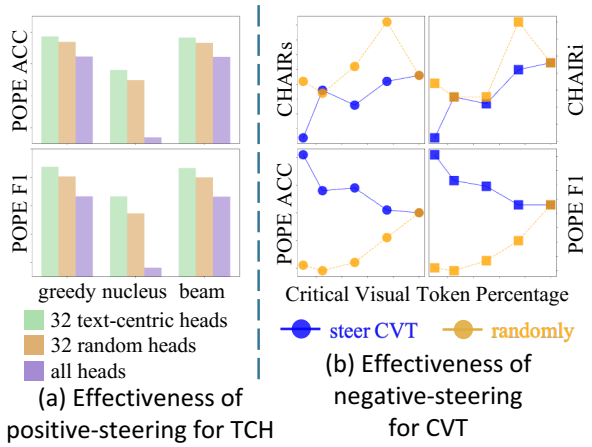

Figure 5: Comparative effectiveness of selective attention steering. (a): Positive steering applied *only* to text-centric heads outperforms random or blanket head selection across various decoding strategies. (b): Negative steering focused on a small subset of critical visual tokens, integrated with contrastive decoding, significantly reduces CHAIR metrics (less hallucination) and boosts POPE scores compared to randomly suppressing visual tokens of the same number.

isons 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.

POPE and CHAIR. Table 1 summarizes both caption-level (CHAIR) and VQA-style (POPE) results. Across every backbone and decoding scheme, ASCD produces the *lowest* CHAIR scores and the *highest* POPE accuracy/F1, outperforming Orig, VCD, and ICD. These gains hold for all three prompt types (random, popular, adversarial; see Appendix "Detailed POPE Results") and remain stable when scaling from 7 B to 13 B or switching to the Qwen-VL architecture, indicating that attention steering mitigates object-

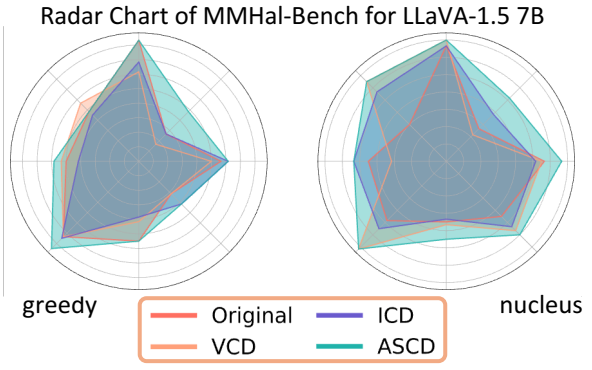

Figure 6: Radar charts of MMHal-Bench results. Each axis represents a different evaluation dimension in MMHal-Bench, and a larger enclosed area indicates better overall performance.

Benchmark	Orig	VCD	ICD	ASCD
MM-VET (↑)	31.2	30.3	33.2	**33.2**
ScienceQA (↑)	67.55	67.55	67.32	**69.51**
GQA (↑)	**61.28**	59.38	59.99	61.27
TextVQA (↑)	57.82	55.07	57.66	**57.91**
MMMU (↑)	0.342	0.333	**0.360**	0.348

Table 2: Scores on five VQA-style benchmarks. The best score in each row is bold. If ASCD ranks second, the best is bold while the ASCD score is underlined.

level hallucinations in a model- and domain-robust manner.

MMHal-Bench. Figure 6 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.

Standard VQA Benchmarks. To verify that ASCD does not sacrifice a model's general visual-question-answering ability, it is evaluated on five widely-used VQA datasets. Across all three representative 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 Table 2.

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

Conclusion

We have shown that existing contrastive methods (e.g., VCD and ICD) inadvertently *shift* the internal attention distribution in MLLMs, prompting us to investigate a more direct and principled way to modulate attention. We propose 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.

Acknowledgments

This work was performed on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research. The authors gratefully acknowledge the scientific support and HPC resources provided by the KIT National High Performance Computing Center (NHR@KIT) under the NHR project 24770, 25312 and 25767.

References

Agrawal, A.; Lu, J.; Antol, S.; Mitchell, M.; Zitnick, C. L.; Batra, D.; and Parikh, D. 2016. VQA: Visual Question Answering. arXiv:1505.00468.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; Ring, R.; Rutherford, E.; Cabi, S.; Han, T.; Gong, Z.; Samangooei, S.; Monteiro, M.; Menick, J.; Borgeaud, S.; Brock, A.; Nematzadeh, A.; Sharifzadeh, S.; Binkowski, M.; Barreira, R.; Vinyals, O.; Zisserman, A.; and Simonyan, K. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. arXiv:2204.14198.

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. arXiv:1707.07998.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Bai, Z.; Wang, P.; Xiao, T.; He, T.; Han, Z.; Zhang, Z.; and Shou, M. Z. 2024. Hallucination of Multimodal Large Language Models: A Survey. arXiv:2404.18930.

Ben-Kish, A.; Yanuka, M.; Alper, M.; Giryes, R.; and Averbuch-Elor, H. 2024. Mitigating Open-Vocabulary Caption Hallucinations. arXiv:2312.03631.

Bi, J.; Wang, Y.; Chen, H.; Xiao, X.; Hecker, A.; Tresp, V.; and Ma, Y. 2024. Visual Instruction Tuning with 500x Fewer Parameters through Modality Linear Representation-Steering. *arXiv preprint arXiv:2412.12359*.

Bi, J.; Wang, Y.; Yan, D.; Xiao, X.; Hecker, A.; Tresp, V.; and Ma, Y. 2025. Prism: Self-pruning intrinsic selection method for training-free multimodal data selection. *arXiv preprint arXiv:2502.12119*.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Chen, H.; Li, H.; Zhang, Y.; Bi, J.; Zhang, G.; Zhang, Y.; Torr, P.; Gu, J.; Krompass, D.; and Tresp, V. 2025a. FedBiP: Heterogeneous One-Shot Federated Learning with Personalized Latent Diffusion Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 30440–30450.

Chen, X.; Wang, X.; Changpinyo, S.; Piergiovanni, A.; Padlewski, P.; Salz, D.; Goodman, S.; Grycner, A.; Mustafa, B.; Beyer, L.; Kolesnikov, A.; Puigcerver, J.; Ding, N.; Rong, K.; Akbari, H.; Mishra, G.; Xue, L.; Thapliyal, A.; Bradbury, J.; Kuo, W.; Seyedhosseini, M.; Jia, C.; Ayan, B. K.; Riquelme, C.; Steiner, A.; Angelova, A.; Zhai, X.; Houlsby, N.; and Soricut, R. 2023a. PaLI: A Jointly-Scaled Multilingual Language-Image Model. arXiv:2209.06794.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; Li, B.; Luo, P.; Lu, T.; Qiao, Y.; and Dai, J. 2024. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. arXiv:2312.14238.

Chen, Z.; Zhao, R.; Luo, C.; Sun, M.; Yu, X.; Kang, Y.; and Huang, R. 2025b. Sifthinker: Spatially-aware image focus for visual reasoning. *arXiv preprint arXiv:2508.06259*.

Chen, Z.; Zhu, Y.; Zhan, Y.; Li, Z.; Zhao, C.; Wang, J.; and Tang, M. 2023b. Mitigating Hallucination in Visual Language Models with Visual Supervision. arXiv:2311.16479.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M. F.; Parikh, D.; and Batra, D. 2017. Visual Dialog. arXiv:1611.08669.

Deng, A.; Chen, Z.; and Hooi, B. 2024. Seeing is Believing: Mitigating Hallucination in Large Vision-Language Models via CLIP-Guided Decoding. arXiv:2402.15300.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 2024a. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. arXiv:2311.17911.

Huang, W.; Liu, H.; Guo, M.; and Gong, N. Z. 2024b. Visual Hallucinations of Multi-modal Large Language Models. arXiv:2402.14683.

Hudson, D. A.; and Manning, C. D. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. arXiv:1902.09506.

Huo, F.; Xu, W.; Zhang, Z.; Wang, H.; Chen, Z.; and Zhao, P. 2024. Self-Introspective Decoding: Alleviating Hallucinations for Large Vision-Language Models. arXiv:2408.02032.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2023. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. arXiv:2311.16922.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023b. Evaluating Object Hallucination in Large Vision-Language Models. arXiv:2305.10355.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved Baselines with Visual Instruction Tuning.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024a. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023b. Visual Instruction Tuning.

Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; and Peng, W. 2024b. A Survey on Hallucination in Large Vision-Language Models. arXiv:2402.00253.

Liu, S.; Zheng, K.; and Chen, W. 2024. Paying More Attention to Image: A Training-Free Method for Alleviating Hallucination in LVLMs. arXiv:2407.21771.

Lovenia, H.; Dai, W.; Cahyawijaya, S.; Ji, Z.; and Fung, P. 2024. Negative Object Presence Evaluation (NOPE) to Measure Object Hallucination in Vision-Language Models. arXiv:2310.05338.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. arXiv:2209.09513.

Lu, S.; Lian, Z.; Zhou, Z.; Zhang, S.; Zhao, C.; and Kong, A. W.-K. 2025. Does FLUX Already Know How to Perform Physically Plausible Image Composition? *arXiv preprint arXiv:2509.21278*.

Lu, S.; Liu, Y.; and Kong, A. W.-K. 2023. Tf-icon: Diffusion-based training-free cross-domain image composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2294–2305.

Lu, S.; Wang, Z.; Li, L.; Liu, Y.; and Kong, A. W.-K. 2024a. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6430–6440.

Lu, S.; Zhou, Z.; Lu, J.; Zhu, Y.; and Kong, A. W.-K. 2024b. Robust watermarking using generative priors against image editing: From benchmarking to advances. *arXiv preprint arXiv:2410.18775*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2023. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2019. Object Hallucination in Image Captioning. arXiv:1809.02156.

Rong, X.; Huang, W.; Liang, J.; Bi, J.; Xiao, X.; Li, Y.; Du, B.; and Ye, M. 2025. Backdoor Cleaning without External Guidance in MLLM Fine-tuning. *arXiv preprint arXiv:2505.16916*.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards VQA Models That Can Read. arXiv:1904.08920.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; Keutzer, K.; and Darrell, T. 2023. Aligning Large Multimodal Models with Factually Augmented RLHF. arXiv:2309.14525.

Team, P.-. 2024a. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219.

Team, Q. 2024b. Qwen2 Technical Report. arXiv:2407.10671.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Wang, X.; Pan, J.; Ding, L.; and Biemann, C. 2024. Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding. arXiv:2403.18715.

Xiao, Z.; Wang, Z.; Ma, W.; Zhang, Y.; Shen, W.; Wang, Y.; Gong, L.; and Liu, Z. 2025. Mitigating Posterior Salience Attenuation in Long-Context LLMs with Positional Contrastive Decoding. arXiv:2506.08371.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2024. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. arXiv:2308.02490.

Yu, X.; Chen, Z.; Zhang, Y.; Lu, S.; Shen, R.; Zhang, J.; Hu, X.; Fu, Y.; and Yan, S. 2025a. Visual Document Understanding and Question Answering: A Multi-Agent Collaboration Framework with Test-Time Scaling. *arXiv preprint arXiv:2508.03404*.

Yu, X.; Chen, Z.; Zhang, Y.; Lu, S.; Shen, R.; Zhang, J.; Hu, X.; Fu, Y.; and Yan, S. 2025b. Visual document understanding and question answering: A multi-agent collaboration framework with test-time scaling. *arXiv preprint arXiv:2508.03404*.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; Wei, C.; Yu, B.; Yuan, R.; Sun, R.; Yin, M.; Zheng, B.; Yang, Z.; Liu, Y.; Huang, W.; Sun, H.; Su, Y.; and Chen, W. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. arXiv:2311.16502.

Zhao, L.; Deng, Y.; Zhang, W.; and Gu, Q. 2024. Mitigating Object Hallucination in Large Vision-Language Models via Classifier-Free Guidance. arXiv:2402.08680.

Zhou, B.; Hu, Y.; Weng, X.; Jia, J.; Luo, J.; Liu, X.; Wu, J.; and Huang, L. 2024. TinyLLaVA: A Framework of Small-scale Large Multimodal Models. arXiv:2402.14289.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. arXiv:2304.10592.

Zhu, L.; Ji, D.; Chen, T.; Xu, P.; Ye, J.; and Liu, J. 2024. IBD: Alleviating Hallucinations in Large Vision-Language Models via Image-Biased Decoding. arXiv:2402.18476.