AHEAD: Attention Head Energy-Aware Dynamics for Hallucination Mitigation in MLLMs

Jiale Chang, Ying Li, Siliang Tang, Yueting Zhuang
Zhejiang University
{jialechang, ying.li, siliang, yzhuang}@zju.edu.cn

Abstract

Multimodal large language models excel at vision-language tasks but remain prone to hallucinations that undermine their reliability. Existing approaches predominantly treat hallucinations as classification errors, overlooking the heterogeneous behaviors of attention heads and their dynamic influences during inference. We revisit MLLM reasoning from an energy perspective and identify that hallucinations stem from imbalances between visual potential and language prior potential: when visual information is ambiguous or language priors dominate, attention heads tend to be driven by linguistic statistical patterns, generating content inconsistent with visual evidence. We propose AHEAD, a framework that quantifies the energetic properties of each attention head during object generation through two potential networks—the Visual Grounding Potential Network and the Language Prior Potential Network—and dynamically adjusts their contributions at inference time. Specifically, we amplify attention heads with strong visual grounding capacity while suppressing those overly reliant on language priors. Experiments across multiple benchmarks demonstrate that AHEAD significantly reduces hallucination rates without fine-tuning the base MLLM while maintaining generation quality.

1 Introduction

Multimodal large language models have demonstrated remarkable capabilities in vision-language tasks, yet hallucinations remain a critical barrier to their reliable deployment in precision-sensitive domains. Hallucinations refer to model-generated content that contradicts visual input, such as describing non-existent objects or incorrect attribute relationships. Despite extensive research efforts toward hallucination mitigation, existing methods predominantly frame the problem as classification errors and intervene through contrastive decoding or logits adjustment, overlooking a fundamental aspect: the heterogeneous behaviors of attention heads and their dynamic influences during inference.

We revisit MLLM reasoning from an energy perspective. Our analysis reveals that hallucinations stem from imbalances between visual potential and language prior potential: when visual information is ambiguous or language priors dominate, the inference trajectory deviates from visual evidence and slides into spurious local minima shaped by linguistic statistical patterns. Using the logit lens technique, we observe that after briefly predicting the correct object in middle layers, models undergo "trajectory escape" (Mir, 2025; Zhang and Zhou, 2025; Ma et al., 2025; Vu et al., 2025) driven by language co-occurrence priors, ultimately outputting hallucinated objects. Further attention analysis demonstrates that when generating hallucinated objects, most attention heads fail to focus on image regions at critical layers, revealing the dynamical process of visual grounding failure competing with language priors.

Building on this insight, we propose **AHEAD**, a framework that quantifies the energetic properties of each attention head during object generation through two potential networks—the Visual Grounding Potential Network and the Language Prior Potential Network—and dynamically adjusts their contributions at inference time. Specifically, we amplify attention heads with strong visual grounding capacity while suppressing those overly reliant on language priors. Experiments across multiple benchmarks demonstrate that **AHEAD** significantly reduces hallucination rates without fine-tuning the base MLLM while maintaining generation quality.

Findings of the Association for Computational Linguistics: ACL 2026, pages 8728–8739

2 Related Work

2.1 Hallucination in Multimodal Large Language Models

The rapid evolution of Multimodal Large Language Models (MLLMs), including LLaVA (Liu et al., 2023b), InstructBLIP (Dai et al., 2023), Qwen-VL (Bai et al., 2023), and others such as InternLM-XComposer (Zhang et al., 2023), DeepSeek-VL (Lu et al., 2024a), and MiniCPM-V (Yao et al., 2024), has revolutionized vision-language understanding. Despite their success, these models suffer significantly from hallucinations (Liu et al., 2024; Bai et al., 2024; Ji et al., 2023), where generated content diverges from visual facts. To quantify this, extensive benchmarks have been proposed, ranging from object existence metrics like CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023b), to comprehensive suites like AMBER (Wang et al., 2023), MME (Fu et al., 2023), HallusionBench (Guan et al., 2024), and the recent CityAnchor (Li et al., 2025). Theoretical analyses attribute these failures to various factors, including "knowledge overshadowing" where linguistic priors suppress visual cues (Zhang et al., 2024), inherent statistical biases in internal states (Orgad et al., 2024; Chen et al., 2024a; Xu et al., 2024), and issues arising from long-context processing (Lu et al., 2024b; Zheng et al., 2025).

2.2 Mitigation Strategies

Current mitigation efforts generally fall into two paradigms: training-time alignment and inference-time intervention.

Training-based Methods. Enhancing data quality and alignment objectives is a primary direction. Reinforcement Learning from Human Feedback (RLHF) has been adapted for MLLMs to penalize hallucinations (Yu et al., 2023b; Sun et al., 2023). More recent approaches leverage Direct Preference Optimization (DPO) and hierarchical preference learning to align models without complex reward modeling (Yang et al., 2025; Fu et al., 2025). Other strategies involve robust instruction tuning (Liu et al., 2023a), generating negative constraints (Yu et al., 2023a), or employing post-hoc correction via external tools (Yin et al., 2023). While effective, these methods incur high retraining costs.

Inference-time Decoding Strategies. Non-training interventions modify the sampling process to reject hallucinations. Contrastive Decoding (CD) (Li et al., 2023a) serves as a foundational technique, inspiring multimodal variants like VCD (Leng et al., 2023), M3ID (Favero et al., 2024), and IAT (Tang et al., 2025), which contrast predictions against distorted inputs or anchor tokens. OPERA (Huang et al., 2024) and DoLa (Chuang et al., 2024) introduce penalties for over-trust and layer-wise contrasts, respectively. Recent advancements in 2025 focus on dynamic correction (Wang et al., 2025), coarse-to-fine feedback (Cao et al., 2025), and multi-path contrastive decoding (Ruan et al., 2025). Furthermore, attention-based interventions have emerged, such as manipulating attention patterns (Zhao et al., 2025; Zhang et al., 2025) or utilizing cross-level trusted interventions (Chen et al., 2024b) to ground generation.

Distinct from global decoding adjustments, our AHEAD framework delves into the fine-grained energy dynamics of attention heads. We identify and rectify the "trajectory escape" phenomenon—where linguistic priors override visual potentials—by dynamically reweighting attention heads during inference.

3 Method

3.1 Inference as Trajectory Evolution on an Energy Manifold

We model the inference process of MLLMs as a dynamical trajectory evolution on a high-dimensional semantic manifold. Specifically, given an input image I and a text prompt, the MLLM progressively generates a response through L transformer layers. At each layer $l \in \{0, 1, \ldots, L-1\}$, the hidden state $\mathbf{h}^{(l)} \in \mathbb{R}^d$ represents a state point on this manifold, where d denotes the hidden dimension. As l increases from 0 to $L-1$, these state points trace out a continuous trajectory $\mathcal{T} = \{\mathbf{h}^{(0)}, \mathbf{h}^{(1)}, \ldots, \mathbf{h}^{(L-1)}\}$ along the manifold.

The evolution of this trajectory is governed by a potential field $\Phi(\mathbf{h}^{(l)})$, which constrains the movement of state points. Drawing on principles from physics, we decompose this potential field into two components:

Visual Potential $\Phi_{\text{visual}}(\mathbf{h}^{(l)}, I)$: This component originates from the input image I and guides the inference trajectory toward regions aligned with visual evidence. We term these regions *low-energy wells*, where the model's judgments about object existence are confident and grounded.

Language Prior Potential $\Phi_{\text{prior}}(\mathbf{h}^{(l)})$: This

component stems from the pre-trained language model's probability distribution P_{LLM}. It attempts to pull the trajectory toward regions maximizing linguistic plausibility, even when such regions contradict visual facts. This reflects the model's internalized co-occurrence statistics learned during pre-training.

3.2 Energy Dissipation and Trajectory Escape

Under this framework, hallucination is no longer viewed as a simple classification error. Instead, we redefine it as a thermodynamic phase transition or trajectory escape phenomenon.

Ideal Scenario. When the model describes an image, the inference trajectory should be strongly captured by Φ_{visual}, maintaining a low-entropy, low-energy stable state. In this regime, the trajectory evolves smoothly with minimal curvature, adhering to the *principle of least action*. The hidden states exhibit low variance, and the output logits distribution remains sharp and concentrated.

Hallucination Scenario. However, when visual information is ambiguous (e.g., small objects, occlusions) or language priors are excessively strong (e.g., adversarial prompts), visual guidance weakens. If the model then over-thinks—performing excessive reasoning—the internal state's energy surges, manifesting as chaos and entropy increase in the logits distribution. To alleviate this cognitive dissonance, the model tends to slide into the nearest deep potential well, typically a spurious local minimum shaped by language priors. This is the underlying physical process of hallucination: the inference trajectory breaks through the visual potential barrier and escapes into the language-prior-dominated region.

Empirical Observations. We conduct two preliminary experiments to validate this hypothesis:

Experiment 1: Trajectory Escape in Middle Layers. Using the logit lens technique, we visualize intermediate predictions at each layer. As shown in Figure 1, when the model attempts to describe a lantern in the image, the hidden state at middle layers (around layer 21) briefly predicts the token "lan" (likely attempting "lantern"). However, this trajectory is subsequently diverted by language priors. Since the language model frequently observed co-occurrences of "clock" and "vase" with "living room" during pre-training, the internalized semantic signal prevails, causing the final output to hallucinate these objects. Notably, even for "book"—an object genuinely present in the image—attention map analysis reveals the model is not attending to the correct regions. Its correct prediction is merely coincidental, underscoring that hallucination stems not only from visual grounding failure but from a competition between visual and language potentials.

Experiment 2: Visual Neglect Across Attention Heads. We compare attention patterns of 32 multi-head self-attention (MHSA) heads across layers for non-hallucinated tokens (e.g., "girl") versus hallucinated tokens (e.g., "clock"). As illustrated in Figure 2, before layer 17, most attention heads generating "clock" fail to attend to image regions. This indicates that during critical middle layers—where visual information should be integrated—the model has already deviated from the visually-grounded trajectory.

Key Insight. These observations reveal that hallucinations are driven by two competing signals: visual potential (weakened when images lack salient features) and language prior potential (strengthened when linguistic patterns dominate). This dual-force framework lays the foundation for our Potential Network modeling in Section 3.3.

3.3 Potential Field Quantification

To operationalize the energy manifold framework, we formulate concrete metrics to quantify the visual potential Φ_{visual} and language prior potential Φ_{prior} introduced in Section 3.1. This quantification enables us to identify attention heads that shape the inference trajectory and assess their contributions to hallucination generation.

3.3.1 Object-level Sample Construction and Attention Representation

Given an input image I and text prompt, the model generates a descriptive text sequence. We extract mentioned objects from the generated text and compare them against the ground-truth annotations of I, partitioning the object set into two categories:

Grounded object set $\mathcal{O}^+$: Objects that exist in the image with annotated bounding boxes.

Hallucinated object set $\mathcal{O}^-$: Objects mentioned in the generated text but absent from the image.

For each object o, we define its first occurrence position in the generated sequence as $t(o)$. All subsequent analyses concerning object o are conducted at this time step $t(o)$.

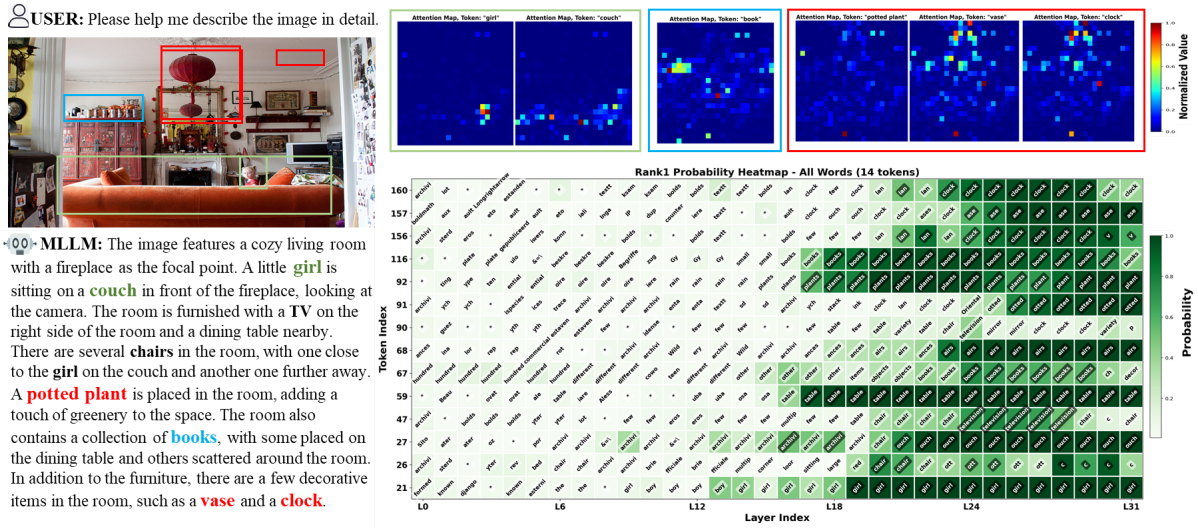

Figure 1: Trajectory escape in middle layers. We visualize the intermediate predictions at each layer of LLaVA-1.5-7B using the logit lens technique.

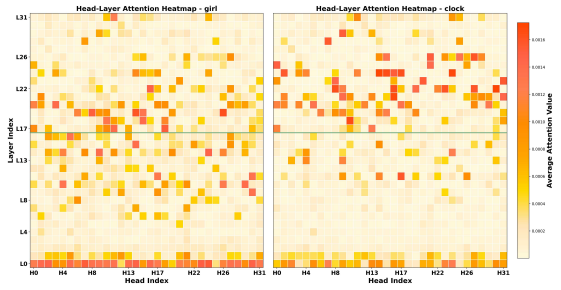

Figure 2: Visual neglect across attention heads during hallucination generation. We compare the layer-wise attention distributions of 32 MHSA heads in LLaVA-1.5-7B when generating the non-hallucinated token "girl" versus the hallucinated token "clock".

Text-to-Vision Attention. At position $t(o)$, any attention head (l, n) in the model produces an attention distribution from all preceding text tokens to all subsequent tokens. The attention weights allocated to visual tokens are denoted as:

$$a_o^{(l,n)}(i), \quad i \in \mathcal{V}, \tag{1}$$

where l indexes the Transformer layer, n indexes the attention head within that layer, $\mathcal{V}$ represents the set of all visual tokens, and $a_o^{(l,n)}(i)$ denotes the attention weight assigned by the (l, n)-th head to the i-th visual token when generating object o.

3.3.2 Visual Potential

The visual potential is employed to assess whether attention tracks properly engage with genuine visual evidence during the inference process, thereby determining whether the model is attracted to the spatial regions aligned with the object being generated. Intuitively, a reliable attention head generating a grounded object should satisfy two criteria: (1) sufficiently utilize visual information with adequate visual attention mass, and (2) its visual attention should be spatially aligned with the object boundary.

Object-level Visual Attraction Score. For a grounded object $o \in \mathcal{O}^+$, we define the object-level visual attraction score for the (l, n)-th attention head as:

$$S_{\text{vis}}^{(l,n)}(o) = m_{\text{img}}^{(l,n)}(o) \cdot \text{IoU}^{(l,n)}(o). \tag{2}$$

Each component is defined below.

(1) Visual Attention Mass. The total attention mass allocated to visual tokens is:

$$m_{\text{img}}^{(l,n)}(o) = \sum_{i \in \mathcal{V}} a_o^{(l,n)}(i). \tag{3}$$

This metric captures the actual proportion of attention distributed to visual modality when generating the object, reflecting the model's reliance on visual information.

(2) Attention Bounding Box Alignment IoU. We first construct an attention-induced spatial region by aggregating regions with attention above a threshold:

$$\mathcal{R}_o^{(l,n)} = \bigcup_{i: a_o^{(l,n)}(i) > \tau} R_i, \tag{4}$$

where R_i denotes the spatial region corresponding to visual token i in the image plane, and τ is a small

threshold to filter out noise attention. Subsequently, we define the Intersection over Union (IoU) between this region and the ground-truth bounding box B_o as:

$$\mathrm{IoU}^{(l,n)}(o) = \frac{\mathrm{Area}\left(\mathcal{R}_o^{(l,n)} \cap B_o\right)}{\mathrm{Area}\left(\mathcal{R}_o^{(l,n)} \cup B_o\right)}. \quad (5)$$

Here, B_o denotes the ground-truth bounding box for object o, $\mathrm{Area}(\cdot)$ computes the region area, and $\mathcal{R}_o^{(l,n)}$ is the attention-induced visual focus region. This metric quantifies whether attention spatially aligns with the complete object region, preventing attention from erroneously concentrating on small local patches within the bounding box and thereby misjudging it as effective visual correspondence.

Visual Potential Definition. Based on the object-level visual attraction score, we define the visual potential for attention head (l, n) as:

$$U_{\mathrm{vis}}(l, n) = -\mathbb{E}_{o \in \mathcal{O}^+}\left[S_{\mathrm{vis}}^{(l,n)}(o)\right]. \quad (6)$$

A lower potential indicates that this attention head is more prone to being attracted by genuine visual evidence during inference, corresponding to a low-energy stable direction in the semantic manifold.

3.3.3 Language Prior Potential

The language prior potential is utilized to assess, when the attention head generates objects, whether it drives the generation of grounded objects toward hallucinated objects, reflecting whether it is dominated by linguistic statistical priors and biased away from visual constraints.

Per-token Log-probability Gain. For an object o, let its corresponding generated token be $w(o)$. At its first occurrence position $t(o)$, we define the log-probability gain contributed by attention head (l, n) to this token as:

$$\Delta^{(l,n)}(w(o)) = \log P\left(w(o) \mid \mathbf{h}_{t(o)} + \mathbf{H}_{t(o)}^{(l,n)}\right) - \log P\left(w(o) \mid \mathbf{h}_{t(o)}\right). \quad (7)$$

Here, $P(\cdot)$ represents the model's output probability distribution under given hidden states, $\mathbf{h}_{t(o)}$ denotes the hidden state from the previous layer (layer $l-1$) at position $t(o)$, and $\mathbf{H}_{t(o)}^{(l,n)}$ is the output of the (l, n)-th attention head at position $t(o)$. This metric quantifies the direct causal contribution of this attention head to object token generation by measuring the probability change when adding its output to the input hidden state.

To stabilize the numerical range, we define:

$$s_{\mathrm{lang}}^{(l,n)}(o) = \tanh\left(\Delta^{(l,n)}(w(o))\right). \quad (8)$$

Language Prior Potential Definition. Combining grounded and hallucinated objects, we define the language prior potential for attention head (l, n) as:

$$U_{\mathrm{lang}}(l, n) = -\mathbb{E}_{o \in \mathcal{O}^+}\left[s_{\mathrm{lang}}^{(l,n)}(o)\right] + \mathbb{E}_{o \in \mathcal{O}^-}\left[\max\left(0, s_{\mathrm{lang}}^{(l,n)}(o)\right)\right]. \quad (9)$$

The first term quantifies the head's promotion strength toward grounded object tokens, while the second term penalizes its positive promotion toward hallucinated object tokens. This potential characterizes the statistical linguistic prior's dominant degree: higher potentials indicate that this attention head is more likely to push generation inconsistent with visual evidence.

The visual potential and language prior potential jointly delineate the physical roles of attention heads in the inference process from spatial correspondence and probability promotion perspectives. Together, they determine the stability or escape tendency of the inference trajectory on the high-dimensional semantic manifold, providing an operational foundation for subsequent energy guidance and hallucination suppression.

3.4 Potential Network Modeling and Energy-Guided Head Reweighting

The preceding analysis reveals that attention heads exhibit heterogeneous behaviors in multimodal inference: some heads stably attract the inference trajectory toward low-energy regions aligned with visual evidence, while others exhibit pronounced language prior tendencies that drive trajectory escape when visual constraints are insufficient. To operationalize this insight, we propose two classes of Potential Networks that model the object-conditioned behavior of attention heads and dynamically reweight them during inference in an energy-guided manner.

3.4.1 Head-level Behavior Representation

For any attention head (l, n), we construct an object-conditioned head behavior representation at position $t(o)$:

$$\phi^{(l,n)}(o) \in \mathbb{R}^d. \quad (10)$$

3.4.2 Visual Grounding Potential Network

The Visual Grounding Potential Network (VGPN) learns the strength of an attention head's capacity to be attracted by genuine visual evidence. Its training objective derives from the object-level visual attraction score $S_{\text{vis}}^{(l,n)}(o)$ defined in Section 3.3, which is computed only through grounded object annotations during training.

VGPN is modeled as a parameter-shared regression network:

$$\hat{s}_{\text{vis}}^{(l,n)}(o) = f_{\text{VGPN}}\left(\phi^{(l,n)}(o)\right), \tag{11}$$

where f_{VGPN} is a lightweight multilayer perceptron shared across all layers and attention heads. During training, supervision is applied only to the grounded object set $\mathcal{O}^+$, with the optimization objective:

$$\mathcal{L}_{\text{vis}} = \mathbb{E}_{o\in\mathcal{O}^+}\left[\ell\left(\hat{s}_{\text{vis}}^{(l,n)}(o), S_{\text{vis}}^{(l,n)}(o)\right)\right], \tag{12}$$

where $\ell(\cdot)$ denotes a regression loss function. From an energy perspective, $\hat{s}_{\text{vis}}^{(l,n)}(o)$ represents the strength of the visual low-energy well formed by the attention head for object o. A higher value indicates that the head is more readily attracted by genuine visual evidence, thereby facilitating stable inference trajectories.

3.4.3 Language Prior Potential Network

The Language Prior Potential Network (LPPN) models the linguistic driving force exerted by attention heads during object generation. Unlike the visual potential, the language potential does not distinguish whether an object is grounded or hallucinated; rather, it characterizes the intrinsic generative tendency driven by the attention head.

Based on the language promotion strength $s_{\text{lang}}^{(l,n)}(o)$ defined in Section 3.3, LPPN outputs an object-conditioned driving force scalar:

$$g^{(l,n)}(o) = f_{\text{LPPN}}\left(\phi^{(l,n)}(o)\right), \tag{13}$$

where f_{LPPN} is a parameter-shared multilayer perceptron.

The corresponding training objective is defined as:

$$\begin{aligned}\mathcal{L}_{\text{lang}} = \mathbb{E}_{o\in\mathcal{O}^+}&\left[\log\left(1+\exp\left(-g^{(l,n)}(o)\right)\right)\right]\\ +\mathbb{E}_{o\in\mathcal{O}^-}&\left[\log\left(1+\exp\left(g^{(l,n)}(o)\right)\right)\right].\end{aligned} \tag{14}$$

It is important to note that the groundedness or hallucination attribute of objects is used only during training to constrain the directional tendency of g. During inference, the model does not explicitly classify object types but relies solely on the driving force predicted by the attention head's current behavior. Physically, $g^{(l,n)}(o)$ describes the language prior force applied by the attention head in the current state: its magnitude reflects the driving strength, while its sign determines the action direction in the energy field.

3.4.4 Energy-Guided Head Reweighting

During inference, VGPN and LPPN dynamically generate energy modulation coefficients based on the head behavior under the current object condition, thereby reweighting multi-head attention. For object o at time step $t(o)$, we define the composite modulation weight for attention head (l,n) as:

$$\lambda^{(l,n)}(t(o)) = 1 + \alpha\hat{s}_{\text{vis}}^{(l,n)}(o) + \beta g^{(l,n)}(o), \tag{15}$$

where α, β are scaling coefficients. To ensure numerical stability, we apply a truncation operation:

$$\lambda^{(l,n)}(t(o)) \leftarrow \text{clip}(\lambda^{(l,n)}(t(o)), \epsilon, 2). \tag{16}$$

This weight is used to reweight attention head outputs:

$$\tilde{H}_{t(o)}^{(l)} = \sum_{n=1}^{N} \lambda^{(l,n)}(t(o)) H_{t(o)}^{(l,n)}. \tag{17}$$

This energy-guided mechanism enables attention head contributions to dynamically vary with the inference state: heads with stronger visual grounding potentials are amplified, deepening the visual low-energy well; heads with excessive language driving forces are suppressed, raising the language potential barrier and preventing trajectory escape in critical middle layers.

4 Experiments

4.1 Experimental Setup

Following DeCo's experimental setup, we test with three decoding strategies: greedy, beam search, and nucleus sampling. For AHEAD, potential networks adopt a two-layer MLP architecture with hidden dimension 256. Hyperparameters are set as $\alpha = 0.5$, $\beta = 0.3$ to control the weights of visual potential and language prior potential. We train potential networks on 2000 random images from COCO

Table 1: Training and inference cost comparison.

Method	Training Cost	Inference Cost	AVG POPE F1
VCD	—	109.6ms (×2.6)	83.1
OPERA	—	89.8ms (×2.1)	85.4
DeCo	—	52.4ms (×1.2)	86.3
POVID	20.0 h	42.6ms (×1.0)	86.2
V-DPO	2.3 h	42.6ms (×1.0)	86.9
AHEAD	**0.2 h**	46.1ms (×1.1)	**87.7**

val2014 using AdamW optimizer with learning rate $1e-4$, batch size 32, for 5 epochs; training completes in approximately 12 minutes on a single A100 GPU, with a total of 68.7M parameters across both networks. Hallucination labels are derived automatically by comparing MLLM-generated object mentions against COCO ground-truth annotations, requiring no manual annotation. We conduct experiments on four mainstream multimodal large language models: InstructBLIP, MiniGPT-4, LLaVA-1.5, and Qwen-VL.

Table 1 compares the training and inference overhead of AHEAD against representative baselines on LLaVA-1.5. AHEAD requires only 0.2 hours of training—100× less than POVID—while achieving the best POPE F1 score of 87.7% with a minimal inference overhead of ×1.1.

4.2 Main Results

Hallucination Mitigation Performance. Table 2 presents hallucination mitigation results on the CHAIR benchmark. AHEAD achieves the best performance across all models and decoding strategies, validating the effectiveness of potential network modeling. Under greedy decoding, AHEAD reduces CHAIR_S from 41.2% to 36.5% on InstructBLIP, a reduction of 4.7 percentage points, and CHAIR_I from 14.4% to 11.2%, a reduction of 3.2 percentage points, improving by 4.7% and 3.2% over the second-best method DeCo, respectively. On LLaVA-1.5, AHEAD achieves 32.8% CHAIR_S and 9.3% CHAIR_I, reducing by 12.2 and 5.4 percentage points compared to the vanilla baseline. Notably, AHEAD exhibits strong compatibility with different decoding strategies. Under beam search, AHEAD consistently outperforms OPERA across all models, achieving CHAIR_S of 29.5% on LLaVA-1.5 compared to OPERA's 44.6%, a significant improvement. Under nucleus sampling, AHEAD maintains its lead, reducing CHAIR_S by 14.8 percentage points on InstructBLIP compared to the vanilla baseline. This consistency across decoding strategies demonstrates that potential networks can accurately identify and dynamically adjust attention head contributions without relying on specific decoding mechanisms.

Table 3 shows results on the POPE benchmark. AHEAD achieves the highest F1 scores across all models, with particularly significant improvements on MiniGPT-4. Under greedy decoding, AHEAD achieves 79.1% F1 on MiniGPT-4, improving by 20.6 percentage points over the vanilla baseline's 58.5% and by 1.7 percentage points over DeCo's 77.4%. On LLaVA-1.5, AHEAD reaches 88.1% F1, improving by 5.9 percentage points over the vanilla baseline's 82.2%. On Qwen-VL, AHEAD achieves 87.8% F1, maintaining optimal performance. These results indicate that potential networks effectively identify attention heads overly reliant on language priors and dynamically suppress their contributions through energy-guided reweighting during inference, preventing trajectory escape phenomena.

Visual Grounding and Hallucination Trade-off. Table 4 presents results on the AMBER benchmark, which evaluates the trade-off between hallucination mitigation and visual grounding capability. AHEAD achieves the lowest CHAIR, Hal., and Cog. metrics across all decoding strategies while maintaining competitive Cover scores. Under greedy decoding, AHEAD achieves 6.2% CHAIR, 48.2% Cover, 26.5% Hal., and 2.5% Cog., losing only 0.7 percentage points in coverage while reducing hallucinations compared to the vanilla baseline. Under beam search, AHEAD's CHAIR drops to 5.8%, Hal. to 23.8%, and Cog. to 2.1%, reducing by 0.6, 3.7, and 0.8 percentage points respectively compared to OPERA. This demonstrates that potential networks enhance attention heads with strong visual grounding capacity, not only reducing hallucinations but also maintaining the model's perception of real objects, avoiding performance degradation caused by over-intervention.

General Capability Preservation. Table 6 presents results on the MMVet benchmark to evaluate whether AHEAD compromises the model's general vision-language capabilities. Results show that AHEAD maintains competitive general performance while reducing hallucinations. On LLaVA-1.5, AHEAD outperforms the vanilla baseline and DeCo across all subskills, achieving a total score of 28.8%, improving by 5.2 percentage points over vanilla's 23.6% and by 0.9 percentage points over DeCo's 27.9%. Particularly on recognition (Rec),

Table 2: **CHAIR hallucination evaluation results**. CHAIR evaluates object hallucination in image captions by comparing generated object mentions with ground-truth labels. CHAIR_S denotes the proportion of captions containing hallucinations, and CHAIR_I denotes the proportion of hallucinated objects among mentioned objects. Lower scores indicate fewer hallucinations.

Decoding	Method	InstructBLIP		MiniGPT-4		LLaVA-1.5		Qwen-VL	
		$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$
Greedy	Vanilla	58.8	23.7	31.8	9.9	45.0	14.7	46.0	12.5
	DoLa	48.4	15.9	32.2	10.0	47.8	13.8	46.8	12.9
	DeCo	41.2	14.4	27.0	8.8	37.8	11.1	42.2	10.7
	AHEAD (ours)	**36.5**	**11.2**	**23.2**	**7.1**	**32.8**	**9.3**	**37.8**	**8.9**
Beam Search	Vanilla	55.6	15.8	30.6	9.5	48.8	13.9	41.8	10.8
	OPERA	46.4	14.2	26.2	9.5	44.6	12.8	34.6	9.5
	DeCo	43.8	12.7	24.8	7.5	33.0	9.7	32.0	8.7
	AHEAD (ours)	**40.2**	**10.8**	**22.5**	**6.0**	**29.5**	**8.0**	**29.2**	**7.2**
Nucleus	Vanilla	54.6	24.8	32.6	10.7	48.8	14.2	49.2	13.1
	VCD	58.0	17.0	33.8	11.1	54.0	16.0	46.4	11.9
	DeCo	43.6	12.9	30.8	9.5	42.8	13.2	43.8	11.8
	AHEAD (ours)	**39.8**	**11.1**	**28.5**	**8.0**	**39.2**	**11.5**	**40.5**	**10.2**

Table 3: **POPE hallucination evaluation results**. POPE assesses hallucination through a binary question-answering format, where the model answers whether a queried object exists in the scene. Performance is measured using the standard F1 score. To evaluate robustness, POPE splits queried objects into three subsets: random (arbitrary samples), popular (frequent objects), and adversarial (visually similar to ground-truth).

Decoding	Method	InstructBLIP	MiniGPT-4	LLaVA-1.5	Qwen-VL
		F1 ↑	F1 ↑	F1 ↑	F1 ↑
Greedy	Vanilla	80.0	58.5	82.2	85.2
	DoLa	83.4	72.8	83.2	85.8
	DeCo	84.9	77.4	86.7	86.3
	AHEAD	**86.2**	**79.1**	**88.1**	**87.8**
Beam Search	Vanilla	84.4	70.3	84.9	85.3
	OPERA	84.8	73.3	85.4	86.1
	DeCo	84.9	77.9	86.7	86.4
	AHEAD	**86.3**	**79.5**	**88.2**	**87.9**
Nucleus	Vanilla	79.8	52.8	83.1	84.5
	VCD	79.9	56.0	83.1	84.7
	DeCo	81.8	63.8	85.4	85.2
	AHEAD	**83.5**	**66.2**	**86.9**	**86.6**

Table 4: Results on the AMBER image caption dataset. AMBER evaluates hallucinations in image captions by assessing hallucinated object rate (CHAIR), coverage of ground-truth objects (Cover), proportion of hallucinated responses (Hal.), and tendency to mention cognitively biased objects (Cog.). Lower scores for CHAIR, Hal., and Cog. indicate fewer hallucinations, while a higher Cover score reflects stronger visual grounding. Results are based on LLaVA-1.5-7b.

Decoding	Method	LLaVA-1.5			
		CHAIR ↓	Cover ↑	Hal ↓	Cog ↓
Greedy	Vanilla	8.2	48.9	34.3	4.0
	DoLa	8.0	**50.8**	37.5	4.3
	DeCo	6.6	47.5	28.1	2.8
	AHEAD	**6.2**	48.2	**26.5**	**2.5**
Beam Search	Vanilla	7.1	**50.7**	32.4	3.8
	OPERA	6.4	49.0	27.5	2.9
	DeCo	6.3	46.8	25.1	2.4
	AHEAD	**5.8**	47.5	**23.8**	**2.1**
Nucleus	Vanilla	10.2	50.2	43.3	4.5
	VCD	9.0	**51.7**	40.2	4.4
	DeCo	8.3	48.0	37.5	3.4
	AHEAD	**7.8**	49.2	**35.2**	**3.1**

OCR, and mathematics (Math) tasks, AHEAD achieves 33.2%, 22.8%, and 12.1%, improving by 4.4, 8.7, and 8.6 percentage points over the vanilla baseline. On Qwen-VL, AHEAD achieves a total score of 47.1%, maintaining balanced performance improvements across all subskills. These results validate the fine-grained nature of potential network modeling: by only adjusting attention heads related to object generation, AHEAD avoids disrupting the model's overall reasoning capability, achieving dual optimization of hallucination mitigation and general capability preservation.

Cross-Domain Generalization. To assess whether AHEAD overfits to COCO, we evaluate zero-shot transfer on A-OKVQA and GQA—benchmarks covering different tasks and visual domains. Following VCD's protocol, bounding boxes are generated by SEEM without any domain-specific fine-tuning. As shown in Table 5, AHEAD consistently outperforms all baselines on both benchmarks across LLaVA-1.5 and Qwen-VL, achieving +5∼6% gains over the vanilla baseline. This confirms that the

Table 5: Zero-shot cross-domain generalization on A-OKVQA and GQA. Networks are trained on COCO only and evaluated directly without domain-specific fine-tuning.

Model	Method	A-OKVQA Acc	A-OKVQA F1	GQA Acc	GQA F1
LLaVA-1.5	Vanilla	78.13	78.10	77.89	78.13
	+VCD	79.99	81.30	80.16	81.67
	+DeCo	80.64	81.87	82.75	82.39
	+AHEAD	**83.27**	**83.08**	**84.38**	**83.60**
Qwen-VL	Vanilla	78.31	79.18	78.15	79.22
	+VCD	80.07	80.96	80.65	81.43
	+DeCo	81.88	81.96	81.81	82.17
	+AHEAD	**84.07**	**83.50**	**84.74**	**84.61**

Table 6: Results on the MMVet benchmark. MMVet evaluates the general capabilities of multimodal large language models across 14 subskills, including OCR, object recognition, spatial relations, and commonsense reasoning. It uses multiple-choice questions on natural images. We follow the official protocol and report overall accuracy. Rec denotes recognition, OCR denotes optical character recognition, Know denotes knowledge, Gen denotes generation, Spat denotes spatial, and Math denotes mathematics.

Model	Method	Rec ↑	OCR ↑	Know ↑	Gen ↑	Spat ↑	Math ↑	Total ↑
LLaVA-1.5	Vanilla	28.8	14.1	15.5	16.4	15.6	3.5	23.6
	DeCo	32.1	21.5	18.6	20.7	23.7	11.2	27.9
	AHEAD	**33.2**	**22.8**	**19.3**	**21.5**	**24.5**	**12.1**	**28.8**
Qwen-VL	Vanilla	**51.8**	35.3	**41.0**	35.6	38.1	19.2	45.7
	DeCo	50.5	38.2	38.8	33.8	41.7	26.5	46.3
	AHEAD	51.2	**39.1**	40.5	**36.2**	**42.5**	**27.8**	**47.1**

Table 7: Hyperparameter sensitivity analysis.

$\alpha \backslash \beta$	0.0	0.2	**0.3**	0.4	0.6
0.0	83.4	84.2	84.5	84.3	83.8
0.4	86.8	87.4	87.6	87.2	86.3
0.5	86.9	87.5	**87.7**	87.3	86.4
0.8	85.2	85.8	85.9	85.5	84.6

learned potential functions capture generalizable visual-semantic alignment properties rather than COCO-specific patterns.

4.3 Sensitivity analysis

Table 7 reports a sensitivity analysis over α and β on POPE F1 (LLaVA-1.5). The optimal configuration (α=0.5, β=0.3) achieves 87.7% F1. Performance remains above 87.0% for $\alpha \in \{0.4, 0.5\}$ and $\beta \in \{0.2, 0.3, 0.4\}$, confirming robustness across a broad range of values. Both components contribute synergistically: setting either to zero degrades performance, validating the complementary roles of visual and language prior potentials. Excessively large values ($\alpha \geq 0.8$ or $\beta \geq 0.6$) lead to degradation, consistent with the risk of over-intervention disrupting the model's existing reasoning capacity.

5 Conclusion

We revisit MLLM reasoning from an energy perspective and propose AHEAD, which quantifies attention head behaviors through Visual Grounding and Language Prior Potential Networks. By identifying the trajectory escape phenomenon, AHEAD dynamically reweights attention heads to amplify visual grounding while suppressing language-driven hallucinations. Experiments demonstrate that AHEAD significantly reduces hallucination rates across multiple benchmarks without fine-tuning the base MLLM, while maintaining generation quality and compatibility with various decoding strategies.

Limitations

Our work has several limitations. First, AHEAD requires training two lightweight potential networks (VGPN and LPPN) on data with object-level annotations. When applied to MLLMs with different architectures or parameter scales, these networks must be retrained, limiting cross-model generalization. Second, AHEAD requires access to internal attention states for head reweighting and is therefore not applicable to closed-source API-only models—a constraint shared by most attention-based inference-time intervention methods. Finally, while AHEAD effectively mitigates hallucinations caused by visual-language imbalance, its effectiveness diminishes when visual signals are severely degraded or entirely absent, as the energy-guided reweighting mechanism inherently relies on meaningful visual input.

Acknowledgments

We sincerely thank the anonymous reviewers for their constructive comments. We also thank Zhejiang University for providing the computational resources that supported this work.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

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

Zongsheng Cao, Yangfan He, Anran Liu, Jun Xie, Zhepeng Wang, and Feng Chen. 2025. Cofi-dec: Hallucination-resistant decoding via coarse-to-fine generative feedback in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025*, pages 10709–10718.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024a. INSIDE: llms' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. 2024b. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2411.15268*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 14303–14312. IEEE.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Jinlan Fu, Shenzhen Huangfu, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, and See-Kiong Ng. 2025. Chip: Cross-modal hierarchical direct preference optimization for multimodal llms. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *CVPR*, abs/2311.17911.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

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. *CoRR*, abs/2311.16922.

Jinpeng Li, Haiping Wang, Jiabin chen, Yuan Liu, Zhiyang Dou, Yuexin Ma, Sibei Yang, Yuan Li, Wenping Wang, Zhen Dong, and Bisheng Yang. 2025. Cityanchor: City-scale 3d visual grounding with multi-modality LLMs. In *The Thirteenth International Conference on Learning Representations*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023a. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12286–12312. Association for Computational Linguistics.

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

models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305. Association for Computational Linguistics.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Aligning large multi-modal model with robust instruction tuning. *CoRR*, abs/2306.14565.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024a. Deepseek-vl: Towards real-world vision-language understanding. *CoRR*, abs/2403.05525.

Taiming Lu, Muhan Gao, Kuai Yu, Adam Byerly, and Daniel Khashabi. 2024b. Insights into llm long-context failures: When transformers know but don't tell. *Preprint*, arXiv:2406.14673.

Huan Ma, Jiadong Pan, Jing Liu, Yan Chen, Joey Tianyi Zhou, Guangyu Wang, Qinghua Hu, Hua Wu, Changqing Zhang, and Haifeng Wang. 2025. Semantic energy: Detecting llm hallucination beyond entropy. *arXiv preprint arXiv:2508.14496*.

Amir Hameed Mir. 2025. The geometry of truth: Layer-wise semantic dynamics for hallucination detection in large language models. *arXiv preprint arXiv:2510.04933*.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2024. Llms know more than they show: On the intrinsic representation of llm hallucinations. *Preprint*, arXiv:2410.02707.

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

Jiacheng Ruan, Zongyun Zhang, Jingsheng Gao, Wenzhen Yuan, Ting Liu, and Yuzhuo Fu. 2025. Mpi-cd: Multi-path information contrastive decoding for mitigating hallucinations in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025*, pages 4251–4260.

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. *CoRR*, abs/2309.14525.

Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. 2025. Intervening anchor token: Decoding strategy in alleviating hallucinations for mllms. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*.

Minh Vu, Brian K Tran, Syed A Shah, Geigh Zollicoffer, Nhat Hoang-Xuan, and Manish Bhattarai. 2025. Hallufield: Detecting llm hallucinations via field-theoretic modeling. *arXiv preprint arXiv:2509.10753*.

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

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *CoRR*, abs/2311.07397.

Ziwei Xu, Sanjay Jain, and Mohan S. Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *CoRR*, abs/2401.11817.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. 2025. Mitigating hallucinations in large vision-language models via DPO: on-policy data hold the key. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025*, pages 10610–10620.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *CoRR*, abs/2310.16045.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2023a. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. *Preprint*, arXiv:2311.13614.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2023b. RLHF-V: towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *CoRR*, abs/2312.00849.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, and 2 others. 2023. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*.

Yudong Zhang, Ruobing Xie, Xingwu Sun, Yiqing Huang, Jiansheng Chen, Zhanhui Kang, Di Wang, and Yu Wang. 2025. Dhcp: Detecting hallucinations by cross-modal attention pattern in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025*, pages 3555–3564.

Yuji Zhang, Sha Li, Jiateng Liu, Pengfei Yu, Yi R. Fung, Jing Li, Manling Li, and Heng Ji. 2024. Knowledge overshadowing causes amalgamated hallucination in large language models. *CoRR*, abs/2407.08039.

Yukun Zhang and Xueqing Zhou. 2025. Pdetransformer: A continuous dynamical systems approach to sequence modeling. *arXiv preprint arXiv:2510.03272*.

Qiyan Zhao, Xiaofeng Zhang, Yiheng Li, Yun Xing, Xiaosong Yuan, Feilong Tang, Sinan Fan, Xuhang Chen, Da-Han Wang, and Xu-Yao Zhang. 2025. Mca-llava: Manhattan causal attention for reducing hallucination in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025*, pages 3981–3990.

Ge Zheng, Jiaye Qian, Jiajin Tang, and Sibei Yang. 2025. Why lvlms are more prone to hallucinations in longer responses: The role of context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4101–4113.