PERCEPTION-COGNITION CALIBRATION FOR MITIGATING MLLM HALLUCINATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) frequently suffer from hallucinations, generating content that is ungrounded in visual reality. We diagnose this pathology as a compound failure: *perceptual deficiency* caused by the misallocation of visual attention, and *cognitive inertia* stemming from an over-reliance on linguistic priors. In this paper, we propose Perception-Cognition Calibration (PECO), a unified, training-free framework that mitigates MLLM hallucinations by synergizing visual sharpening with prior suppression. Specifically, to counteract perceptual deficiency, we introduce Spatio-Temporal Attention Rectification. This mechanism progressively reclaims the attention budget wasted on uninformative background artifacts (i.e., sink tokens) and redirects it to semantically critical regions, reinforcing fine-grained visual grounding. Furthermore, to overcome cognitive inertia, we devise Sensory-Deprived Contrastive Decoding. By splitting the inference flow into a visually enhanced stream and a sensory-deprived stream, we dynamically penalize hallucination-prone tokens driven by statistical language biases. Extensive evaluations across multiple benchmarks demonstrate that PECO significantly outperforms existing methods, effectively mitigating hallucinations while enhancing general multimodal reasoning capabilities.

1 INTRODUCTION

"Thoughts without content are empty, intuitions without concepts are blind."
— Immanuel Kant

Multimodal Large Language Models (MLLMs) (Wu et al., 2023; Liang et al., 2024), exemplified by LLaVA (Liu et al., 2023b) and Qwen-VL (Bai et al., 2023), have fundamentally reshaped the landscape of general-purpose artificial intelligence. By bridging a vision encoder (e.g., CLIP (Radford et al., 2021)) with an LLM via a learnable connector, these models project rich visual features into the LLM's semantic space, effectively anchoring the generation process to visual evidence (Kang et al., 2025). Such architectural synergy empowers MLLMs to transcend the boundaries of traditional captioning tasks (Mokady et al., 2021), catalyzing unprecedented capabilities in complex visual reasoning (Wang et al., 2024b) and nuanced, human-aligned interactions (Caffagni et al., 2024).

Despite these advancements, MLLMs remain prone to hallucinations (Bai et al., 2024; Huang et al., 2024b), frequently generating content inconsistent with visual evidence—such as fabricating absent objects or misinterpreting scene details (Zhang et al., 2025a; Zhu et al., 2025). This issue severely undermines model reliability, creating a major bottleneck for their deployment in safety-critical fields such as clinical diagnosis (Zhang et al., 2025b) and autonomous driving (Hu et al., 2025).

To address this challenge, recent research has explored various mitigation strategies, ranging from resource-intensive fine-tuning (Sun et al., 2024b; Gunjal et al., 2024) and high-latency post-hoc rectification (Yin et al., 2024; Zhou et al., 2023) to decoding-centric inference-time interventions (Leng et al., 2024; Kang et al., 2025; Wang et al., 2026). Among these, inference-time interventions have garnered substantial attention due to their training-free nature. However, existing methods predominantly adopt a fragmented perspective—they tend to focus exclusively on either enhancing visual perception (boosting the signal) or constraining linguistic biases (dampening the noise). We argue that such isolated treatments are suboptimal; *generation fidelity relies on optimizing the Signal-to-Noise Ratio, which necessitates addressing visual grounding and language priors in tandem.*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this paper, we propose Perception-Cognition Calibration (PECO), a unified, training-free framework that mitigates MLLM hallucinations by synergizing visual sharpening with prior suppression. Specifically, PECO performs dual-front interventions: 1) Perception Enhancement: We introduce Spatio-Temporal Attention Rectification, which progressively reallocates the attention budget from uninformative sink tokens to semantically critical regions, thereby reinforcing fine-grained visual grounding. 2) Cognition Calibration: We design Sensory-Deprived Contrastive Decoding, which contrasts the visually enhanced stream against an internally constructed, sensory-deprived counterpart. This strategy effectively isolates and penalizes hallucination-prone tokens driven solely by linguistic priors, thereby overcoming cognitive inertia. Our main contributions are summarized as follows:

- We provide a root-cause analysis of MLLM hallucinations, attributing this pathology to a compound failure: visual perceptual deficiency and linguistic cognitive inertia.
- We propose PECO, a training-free framework that integrates Spatio-Temporal Attention Rectification and Sensory-Deprived Contrastive Decoding to mitigate MLLM hallucinations.
- Extensive evaluations demonstrate that PECO significantly outperforms existing methods on both visual hallucination tasks and general vision-language tasks.

2 Related Work

2.1 Multimodal Large Language Models

Building upon the advanced reasoning capabilities of LLMs (Bai et al., 2023; Liu et al., 2024a), Multimodal Large Language Models (MLLMs) have demonstrated exceptional proficiency in complex visual understanding. Early frameworks, such as LLaVA (Liu et al., 2023b) and MiniGPT-4 (Zhu et al., 2023), established this paradigm by employing lightweight projectors to align visual representations with the LLM's semantic space. Driven by the demand for fine-grained perception, subsequent research has focused on enhancing input resolution and visual encoding. For instance, Qwen-VL (Wang et al., 2024a) and LLaVA-NeXT (Li et al., 2024) adopt dynamic resolution strategies to capture intricate details, whereas InstructBLIP (Dai et al., 2023) utilizes Q-Formers for spatially-aware feature extraction. Despite these architectural strides, MLLMs are still prone to generating content ungrounded in visual inputs, underscoring the critical need for robust mitigation mechanisms.

2.2 Hallucination Mitigation in MLLMs

Existing strategies for hallucination mitigation can be broadly classified into three primary paradigms.

Training-Based Alignment. A fundamental approach improves visual grounding by fine-tuning models on carefully curated datasets. For instance, methods like LLaVA-RLHF (Sun et al., 2024b) and HA-DPO (Zhao et al., 2023) leverage Reinforcement Learning from Human Feedback and Direct Preference Optimization to penalize hallucinatory generations. Similarly, LRV-Instruction (Liu et al., 2023a) utilizes positive-negative instruction pairs to sharpen object existence discrimination. Despite their efficacy, these methods suffer from massive computational overhead and demand extensive human-annotated preference data, which severely bottlenecks their scalability and transferability.

Post-Hoc Rectification. This paradigm treats the MLLM as a black box and mitigates hallucinations by refining generated content. Frameworks like Woodpecker (Yin et al., 2024) and LURE (Zhou et al., 2023) adopt a generate-then-revise pipeline, utilizing external object detectors (e.g., Grounding DINO (Liu et al., 2024b)) or auxiliary LLMs to correct inaccurate descriptions. However, their multi-stage verification incurs prohibitive latency, limiting their applicability to real-time scenarios.

Inference-Time Interventions. To avoid training costs and post-hoc verification latency, many methods intervene directly during decoding. Contrastive decoding methods such as VCD (Leng et al., 2024) mitigate hallucinations by contrasting logits derived from original and distorted visual inputs. However, such coarse-grained perturbations inevitably introduce noise and may impair general reasoning abilities (Huo et al., 2025). Alternatively, attention-based methods such as VAR (Kang et al., 2025) enhance visual perception by reallocating attention, yet their naive redistribution fails to recover fine-grained semantics. More importantly, existing methods address perceptual and cognitive failures in isolation, overlooking their interplay. In contrast, PECO jointly addresses both through a unified internal calibration framework, effectively mitigating hallucinations.

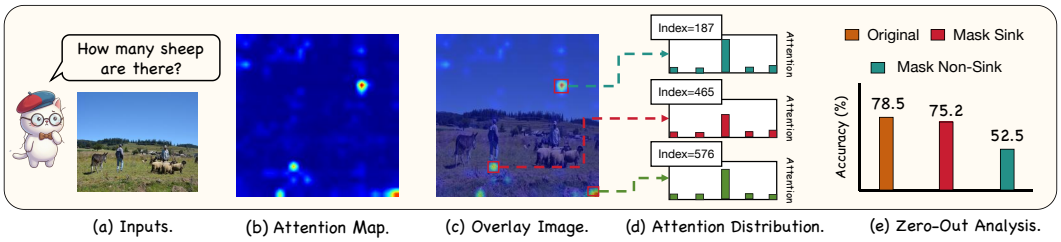


Figure 1: Analysis of visual attention misallocation (LLaVA-1.5-7B). (a) Model inputs. (b) and (c) Averaged attention heatmaps and their overlays on the image. (d) Attention weights of sink tokens. (e) Performance of masking sink versus non-sink visual tokens on VQA^{v2} (Goyal et al., 2017).

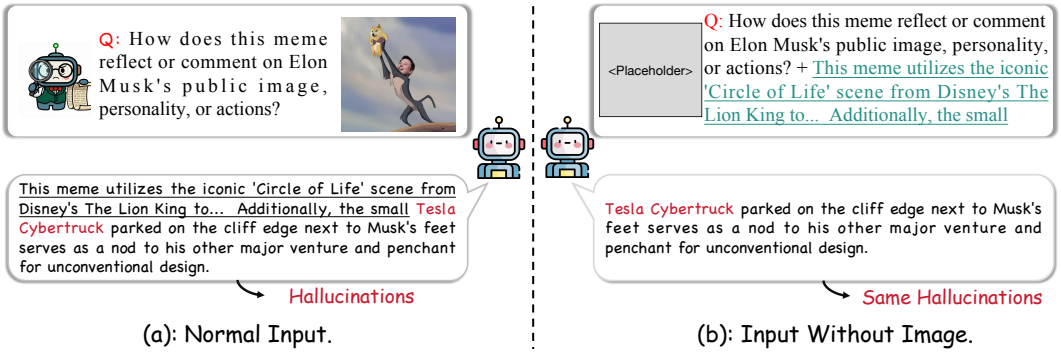


Figure 2: Illustration of cognitive inertia in LLaVA-1.5-13B. (a) Given multimodal inputs, the model fabricates a non-existent object (highlighted in red). (b) Under sensory deprivation, with visual inputs completely removed, the model reproduces the same hallucination based solely on the context.

3 Diagnosing the Roots of Hallucination

In this section, we dissect the fundamental mechanisms underlying MLLM hallucinations.

3.1 The Perceptual Flaw: Visual Attention Misallocation

Ideally, attention serves as a semantic spotlight, grounding generation in relevant visual regions. However, visualizing the attention maps of LLaVA-1.5-7B (Liu et al., 2023b) (Figure 1(b)-(d)) reveals a stark architectural anomaly: rather than aligning with visual saliency, the model disproportionately squanders its attention on uninformative background patches, termed **visual sink tokens** (Kang et al., 2025), whereas genuine semantic targets receive negligible attention. This raises a critical question: *Do these high-attention sink tokens encode essential global context, or are they merely computational artifacts?* We investigate this through a zero-out ablation. If attention magnitude reliably reflects information density, suppressing these sink tokens should substantially degrade model performance.

Diagnosis I: Perceptual Deficiency. As shown in Figure 1(e), masking sink tokens causes only a 3.3% performance drop, whereas masking non-sink tokens leads to a dramatic 26% decline, indicating that sink tokens contain little task-relevant semantic information. This stark contrast reveals severe attention misallocation: meaningful visual signals are overshadowed by background noise, weakening visual grounding and ultimately inducing hallucinations.

3.2 The Cognitive Bias: Linguistic Prior Inertia

While §3.1 exposes the perceptual failure of the multimodal "eye," an equally severe pathology afflicts the reasoning "mind." Many hallucinations arise not from visual misinterpretation, but from excessive reliance on the LLM's linguistic priors. To isolate this effect, we conduct a *sensory deprivation* experiment. Specifically, we identify instances where the MLLM fabricates an object under multimodal inputs. We then strip the model of its visual inputs, feeding it only the historical textual context preceding the hallucination to see if the error persists without any visual stimuli.

DIAGNOSIS II: COGNITIVE INERTIA. As shown in Figure 2, the model reproduces the same hallucinations even without visual input. The persistent fabrication of a "Tesla Cybertruck," conjured purely by the textual trigger "Elon Musk," suggests that multimodal inference devolves into text-driven auto-completion. Rather than grounding its predictions in visual evidence, the model defaults to statistical guesswork. We term this behavior *cognitive inertia*: an entrenched generative bias in which linguistic priors override visual grounding and induce hallucinations.

4 OUR METHOD: PECO

Motivated by these observations, we introduce PECO, which orchestrates a synergy between perception enhancement (§4.2) and cognition calibration (§4.3) to mitigate MLLM hallucinations.

4.1 PRELIMINARIES

Sink Tokens. Recent studies (Sun et al., 2024a) reveal that certain tokens exhibit massive activation outliers across specific feature dimensions. These computational artifacts, termed *sink tokens*, absorb a disproportionate amount of attention mass despite encoding negligible semantic information. Following VAR (Kang et al., 2025), for a given token $\mathbf{x} \in \mathbb{R}^D$, we quantify its sink score $\phi(\mathbf{x})$ based on activation magnitudes within these outlier dimensions $\mathcal{D}_{sink}$:

$$\phi(\mathbf{x}) = \max_{d \in \mathcal{D}_{sink}} \left| \frac{\mathbf{x}[d]}{\sqrt{\frac{1}{D}\sum_{k=1}^{D} \mathbf{x}[k]^2}} \right|, \tag{1}$$

where $\mathbf{x}[d]$ denotes the activation value at the d-th dimension, and $\mathcal{D}_{sink}$ represents the pre-identified set of outlier dimensions specific to each MLLM family (detailed in Appendix A). Leveraging this metric, we formally define the set of sink tokens $\check{\mathcal{I}}^l$ at layer l by applying a threshold κ:

$$\check{\mathcal{I}}^l = \{ j \in \mathcal{I} \mid \phi(\mathbf{x}_j^{l-1}) \geq \kappa \}. \tag{2}$$

Here, $\mathbf{x}_j^{l-1}$ denotes the input hidden state of the j-th token. We distinguish the subset of visual sink tokens as $\check{\mathcal{I}}_{vis}^l$, leaving $\mathcal{I}_{vis} \setminus \check{\mathcal{I}}_{vis}^l$ to represent the informative visual non-sink tokens.

Visual Non-Sink Ratio. To evaluate how effectively each head captures visual semantics, we introduce the *visual non-sink ratio*. This metric quantifies the proportion of the visual attention mass allocated to genuinely informative tokens. Formally, for head h in layer l, it is defined as:

$$r^{l,h} = \sum_{i \in \mathcal{I}_{txt}} \frac{\sum_{j \in \mathcal{I}_{vis} \setminus \check{\mathcal{I}}_{vis}^l} \alpha_{i,j}^{l,h}}{\sum_{j \in \mathcal{I}_{vis}} \alpha_{i,j}^{l,h}}, \tag{3}$$

where $\mathcal{I}_{txt}$ and $\mathcal{I}_{vis}$ denote the index sets of text and visual tokens, respectively, and $\alpha_{i,j}^{l,h}$ represents the attention weight assigned by text token i to visual token j. A higher $r^{l,h}$ signifies that the head actively grounds its representations in meaningful visual content.

Image-Centric Head. Based on the non-sink ratio, we define image-centric heads $\mathcal{H}_{vis}$ as $r^{l,h} \geq \lambda$.

4.2 PERCEPTION ENHANCEMENT

To enhance visual perception, we surgically intervene in image-centric heads $\mathcal{H}_{vis}$. This involves attention reclamation and attention rectification, as shown in Figure 3.

Step 1: Attention Reclamation. For each head $h \in \mathcal{H}_{vis}$, we first quantify the attention mass wasted on non-informative sink tokens $\check{\mathcal{I}}$. We calculate the reclaimable attention budget Ω as (omitting layer l and head h indices for notational brevity):

$$\Omega = \sum_{j \in \check{\mathcal{I}}} \gamma \cdot \alpha_{i,j}, \tag{4}$$

where γ denotes the attention reclamation strength. Correspondingly, sink token attention is attenuated to $\check{\alpha}_{i,j} = (1-\gamma) \cdot \alpha_{i,j}$, effectively harvesting the attention mass Ω for redistribution.

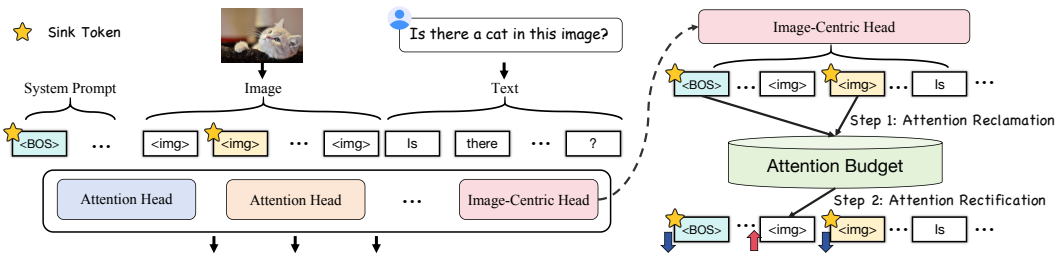


Figure 3: Workflow of perception enhancement, consisting of attention reclamation and rectification.

Step 2: Attention Rectification. With Ω reclaimed, the challenge shifts to where and how to reinvest it. Existing methods naively redistribute this budget to non-sink tokens in proportion to their original attention scores. However, the presence of sink tokens implies that the raw attention distribution is already distorted. To address this, we introduce Spatio-Temporal Attention Rectification (STAR), which calibrates redistribution along both spatial and temporal dimensions.

i) Spatial Dimension: Continuity Rescues Semantics. Anchored in the spatial continuity of objects, visual semantics manifest as contiguous clusters rather than isolated patches. Consequently, a seemingly low-scoring token warrants up-weighting if embedded within a salient neighborhood. To operationalize this, we introduce Spatially-Aware Importance Smoothing (Figure 4). By reshaping visual attention into a 2D spatial map $\mathbf{A}_{vis}$ and applying a smoothing kernel (e.g., 3×3 AvgPool), we prioritize robust regional consensus over noisy individual scores, deriving a contextualized importance map $\mathbf{A}'_{vis}$:

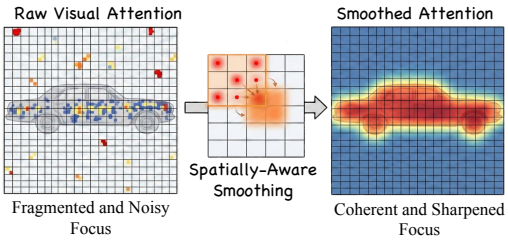


Figure 4: Illustration of the Spatially-Aware Importance Smoothing mechanism.

$$\mathbf{A}'_{vis} = \text{AvgPool}(\mathbf{A}_{vis}). \tag{5}$$

We then redistribute Ω proportionally to this smoothed map $\mathbf{A}'_{vis}$ rather than the noisy map $\mathbf{A}_{vis}$:

$$\tilde{\alpha}_{i,j} = \alpha_{i,j} + \Omega \cdot \frac{\alpha'_{i,j}}{\sum_{k \in \mathcal{I}_{vis} \setminus \check{\mathcal{I}}_{vis}} \alpha'_{i,k}}, \forall j \in \mathcal{I}_{vis} \setminus \check{\mathcal{I}}_{vis}. \tag{6}$$

This strategy effectively ensures the spatial coherence of the enhanced attention.

ii) Temporal Dimension: Combating Signal Decay. Static calibration overlooks visual attention fading, where the model's focus on visual tokens progressively deteriorates as generation lengthens. Therefore, a fixed reclamation ratio is inadequate. To address this, we formulate a Chronological Compensation Function $\gamma(t)$ that modulates the reclamation intensity based on the decoding step t:

$$\gamma(t) = \gamma_{base} + \eta \cdot \left(1 - e^{-\frac{t}{\tau}}\right). \tag{7}$$

Here, the time-varying penalty $\eta \cdot (1 - e^{-t/\tau})$ dynamically suppresses sink tokens as generation advances. The reclaimed budget $\Omega(t)$ is then reinvested via the smoothed map $\mathbf{A}'_{vis}$, forging an attention landscape that is spatially coherent and temporally resilient.

4.3 Cognition Calibration

Having sharpened the "eye," we now calibrate the reasoning "mind." Existing methods approximate language priors through external perturbations, which are coarse-grained and may introduce uncontrolled noise. To overcome this, we propose Sensory-Deprived Contrastive Decoding (SDCD), which mines purely internal signals to precisely disentangle linguistic priors.

Our rationale posits that when deprived of visual perception, the model inevitably resorts to linguistic priors. Therefore, starting from an anchor layer L_{anchor}, we create a parallel stream where informative non-sink tokens within image-centric heads $\mathcal{H}_{vis}$ are surgically ablated. Forced to predict using

solely artifactual noise ($\check{\mathcal{I}}_{vis}$) and textual context, this stream explicitly isolates the model's cognitive inertia. Formally, we achieve this via a binary mask $\mathbf{M}$ that exclusively retains the sink tokens:

$$\mathbf{M}_{i,j} = \begin{cases} 1 & \text{if } j \in \check{\mathcal{I}}_{vis} \\ 0 & \text{if } j \in \mathcal{I}_{vis} \setminus \check{\mathcal{I}}_{vis} \end{cases} \tag{8}$$

We then modulate the attention in the contrastive stream as $\mathbf{A}^c_{vis} = \mathbf{A}_{vis} \odot \mathbf{M}$. This stream yields negative logits, encapsulating pure language priors ungrounded in visual reality.

Contrastive Decoding. Ultimately, we mitigate cognitive inertia by contrasting the visually sharpened stream (V^+) against this isolated language prior (V^-) (Figure 5):

$$P_{\text{cd}}(y_t) = \text{softmax}\Big[(1+\beta)\text{logit}_\theta(y_t|S, V^+, T, y_{<t}) - \beta\text{logit}_\theta(y_t|S, V^-, T, y_{<t})\Big], \tag{9}$$

where β controls the suppression strength, S denotes the system prompt, and T represents the textual input. This formulation penalizes tokens driven by pure linguistic guessing (high in V^-) while reinforcing those grounded in visual evidence (high in V^+). This synergy ensures the MLLM simultaneously *sees clearly* and *thinks critically*.

Adaptive Plausibility Constraint. To prevent the spurious promotion of implausible tokens during contrastive decoding, we impose an adaptive plausibility constraint. Using the enhanced stream as a reliable anchor, we restrict the output to a candidate set:

$$\mathcal{V}_{\text{cand}}(y_{<t}) = \big\{y_t \in \mathcal{V} : P_\theta(y_t|S, V^+, T, y_{<t}) \geq \delta \cdot \max_w P_\theta(w|S, V^+, T, y_{<t})\big\},$$
$$P_{\text{cd}}(y_t) = 0, \quad \text{if } y_t \notin \mathcal{V}_{\text{cand}}, \tag{10}$$

where $\delta \in [0, 1]$ is an adaptive threshold. Larger values impose stricter truncation, encouraging the generated output to remain grounded in visual evidence.

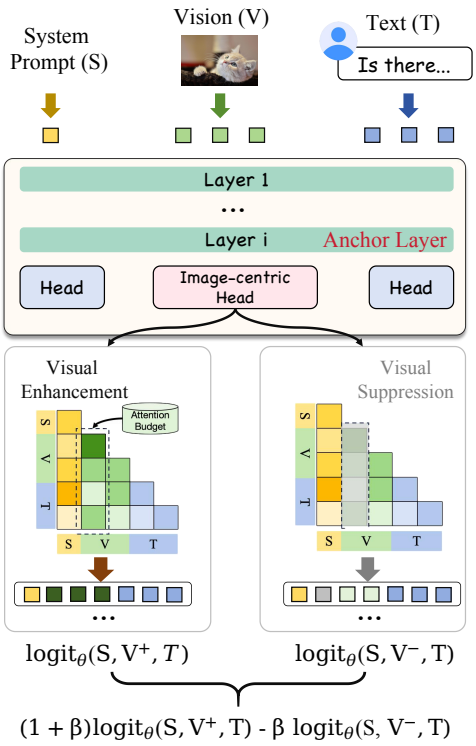


Figure 5: Illustration of the Sensory-Deprived Contrastive Decoding.

5 Experiments

5.1 Experimental Setup

Evaluation Benchmarks. We evaluate PeCo across a suite of benchmarks: 1) **Visual Hallucination Tasks**: We utilize POPE (Li et al., 2023b) and CHAIR (Rohrbach et al., 2018) to assess object existence and generative faithfulness, respectively, alongside MME (Fu et al., 2026) for fine-grained perception. 2) **General Vision-Language Tasks**: To ensure robust general multimodal reasoning capabilities, we further evaluate on MM-Vet (Yu et al., 2023), VizWiz (Gurari et al., 2018), SEED (Li et al., 2023a), GQA (Hudson & Manning, 2019), and MMB (Liu et al., 2024c).

Models. We evaluate PeCo across a diverse range of MLLMs spanning different families: LLaVA-1.5 (Liu et al., 2023b), Qwen2-VL (Wang et al., 2024a), Qwen-VL-Chat (Bai et al., 2023), Qwen3-VL (Bai et al., 2025), MiniGPT-4 (Zhu et al., 2023), InstructBLIP (Dai et al., 2023), Mini-Gemini (Li et al., 2025), and InternVL2 (Chen et al., 2024b). We deploy the 8B variants for Mini-Gemini, InternVL2, and Qwen3-VL, while using the 7B variants for all remaining models.

Baselines. We benchmark PeCo against representative methods across two distinct categories: 1) **Decoding Strategies**, which encompass sampling-based techniques such as DoLa (Chuang et al., 2024), VCD (Leng et al., 2024), HALC (Chen et al., 2024a), RITUAL (Woo et al., 2024), and

Table 1: Performance evaluation of PECO on visual hallucination tasks. The best and second-best results are highlighted in **bold** and underlined, respectively. The model is LLaVA-1.5-7B.

Method	POPE		CHAIR			MME				
	F1↑	Acc.↑	$CHAIR_S$↓	$CHAIR_I$↓	Length	Exist.↑	Count↑	Pos.↑	Color↑	Total↑
Beam Search	85.4	84.0	51.0	15.2	102.2	175.67	124.67	114.00	151.00	565.34
DoLa (Chuang et al., 2024)	80.2	83.1	57.0	15.2	97.5	180.10	127.40	119.30	154.60	594.10
VCD (Leng et al., 2024)	85.3	85.0	51.0	14.9	101.9	184.66	137.33	128.67	153.00	603.66
HALC (Chen et al., 2024a)	83.9	84.0	50.2	12.4	97.2	190.00	143.30	128.30	160.00	621.60
RITUAL (Woo et al., 2024)	85.2	84.3	45.2	13.2	99.2	187.50	139.58	125.00	**164.17**	616.25
SID (Huo et al., 2025)	85.6	85.8	44.2	12.2	99.4	183.90	132.20	127.80	155.90	599.80
OPERA (Huang et al., 2024a)	84.2	85.2	47.0	14.6	95.3	180.67	133.33	111.67	123.33	549.00
DOPRA (Wei & Zhang, 2024)	84.6	84.3	46.3	13.8	96.1	185.67	138.33	120.67	133.00	577.67
AGLA (An et al., 2025)	84.6	85.5	43.0	14.1	98.8	**195.00**	153.89	129.44	156.67	635.00
TAME (Tang et al., 2025)	85.4	85.7	41.3	12.2	98.8	183.00	137.33	129.00	154.67	604.00
VAR (Kang et al., 2025)	86.0	86.5	52.4	14.5	113.0	190.00	138.33	**148.33**	155.00	631.33
EVAS (Zhang et al., 2025a)	85.7	86.0	36.4	9.9	97.7	190.00	148.33	128.00	160.33	626.66
PECO	**86.9**	**87.2**	**33.3**	**8.0**	99.1	193.00	**155.00**	135.67	160.00	**643.67**

SID (Huo et al., 2025), alongside penalty-driven mechanisms like OPERA (Huang et al., 2024a) and DOPRA (Wei & Zhang, 2024); 2) **Attention Interventions**, which directly modulate internal attention landscapes to mitigate hallucinations, including AGLA (An et al., 2025), TAME (Tang et al., 2025), VAR (Kang et al., 2025), and EVAS (Zhang et al., 2025a).

Implementation Details. Main experiments are conducted on NVIDIA A100 GPUs. Following VAR, we set $\kappa = 20$ to identify sink tokens and $\lambda = 0.5$ to isolate image-centric heads. For perception enhancement, spatial smoothing employs a 5×5 average pooling kernel, while attention reclamation uses a base intensity of $\gamma_{base} = 0.5$, with temporal modulation governed by $\eta = 0.55$ and $\tau = 2$. For cognition calibration, we designate the third layer as the anchor L_{anchor}, set the contrastive strength to $\beta = 0.3$, and apply an adaptive threshold $\delta = 0.1$ to truncate low-probability candidates.

5.2 MAIN RESULTS

Performance on Visual Hallucination Tasks. As shown in Table 1, PECO performs strongly across diverse hallucination benchmarks. On POPE, PECO achieves the highest F1 (86.9%) and accuracy (87.2%). Unlike VAR, which performs competitively on POPE but degrades in open-ended generation ($CHAIR_S$: 52.4%), PECO reduces $CHAIR_S$ to 33.3% and $CHAIR_I$ to 8.0% while maintaining an average response length of 99.1, indicating that the improvement does not rely on output truncation. Furthermore, PECO also achieves the highest MME score (643.67), demonstrating strong fine-grained perception. Overall, these results confirm that PECO effectively mitigates visual hallucinations.

Performance on General Vision-Language Tasks. Beyond mitigating hallucinations, it is imperative to ensure that the intervention does not compromise multimodal reasoning capabilities. As evidenced in Table 2, PECO consistently enhances performance compared to the base model across all five benchmarks, achieving accuracy gains of 0.9%, 5.1%, 3.3%, 2.1%, and 2.5%, respectively. Furthermore, PECO significantly outperforms existing methods on VizWiz, SEED, GQA, and MMB. Notably, while several approaches (e.g., DoLa, VCD, and OPERA) exhibit varying degrees of performance degradation, our method avoids this trade-off. These results confirm that PECO effectively mitigates hallucinations while simultaneously improving general multimodal reasoning capabilities.

Table 2: Performance evaluation of PECO on general vision-language benchmarks (LLaVA-1.5-7B).

Method	MM-Vet↑	VizWiz↑	SEED↑	GQA↑	MMB↑
Base	31.1	50.1	57.6	62.0	64.2
DoLa	31.8	48.4	57.2	61.3	63.8
VCD	29.4	50.5	58.3	61.6	61.4
HALC	30.5	50.3	58.0	61.7	63.5
RITUAL	30.6	50.0	58.4	61.5	63.8
SID	31.2	50.8	58.9	62.1	65.0
OPERA	30.0	52.4	59.4	62.0	64.8
DOPRA	30.7	51.8	59.0	62.4	64.7
AGLA	31.5	51.7	60.1	62.2	65.2
TAME	30.5	51.6	59.4	61.7	65.3
VAR	**33.7**	53.7	60.7	63.5	65.1
EVAS	31.7	53.9	60.2	62.3	65.8
PECO	32.0	**55.2**	**60.9**	**64.1**	**66.7**

5.3 MORE ANALYSIS

Analysis of Attention Reclamation Intensity. We then investigate the influence of the initial attention reclamation intensity γ_{base}, which controls the magnitude of the attention budget reclaimed from sink

Table 3: Impact of varying the initial attention reclamation intensity γ_{base} on LLaVA-1.5-7B.

γ_{base}	POPE	CHAIR		MME
	Acc.↑	CHAIR_S↓	CHAIR_I↓	Total↑
0.1	85.4	35.8	9.5	608.30
0.3	86.4	34.2	8.6	629.45
0.5	**87.2**	**33.3**	**8.0**	**643.67**
0.7	86.7	33.9	8.4	632.18
0.9	85.8	35.1	9.2	615.50

Table 4: Impact of varying the cognition calibration strength β on LLaVA-1.5-7B.

β	POPE	CHAIR		MME
	Acc.↑	CHAIR_S↓	CHAIR_I↓	Total↑
0.1	86.2	34.8	8.9	624.15
0.2	86.8	33.9	8.3	634.50
0.3	**87.2**	**33.3**	**8.0**	**643.67**
0.4	86.5	34.2	8.6	629.80
0.5	85.9	35.3	9.2	619.50

tokens. As shown in Table 3, increasing γ_{base} from 0.1 to 0.5 yields consistent improvements across all metrics. For example, POPE accuracy climbs from 85.4% to 87.2%, and the MME total score rises from 608.30 to 643.67. These improvements confirm that enhancing visual perception effectively restores fine-grained details. However, performance plateaus and eventually degrades when γ_{base} exceeds 0.5, with POPE accuracy dropping to 85.8% at $\gamma_{base} = 0.9$. This decline suggests that an overly aggressive reclamation intensity degrades the model's logical reasoning capabilities.

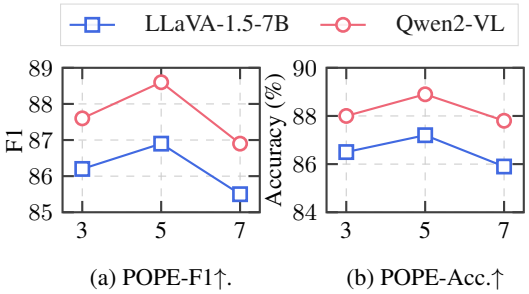
Analysis of Cognition Calibration Strength. We further investigate the impact of the cognition calibration strength β. As detailed in Table 4, model performance improves consistently as β increases from 0.1 to 0.3, peaking at $\beta = 0.3$. These results confirm that our Sensory-Deprived Contrastive Decoding effectively neutralizes cognitive inertia by penalizing tokens driven primarily by language priors. However, increasing β beyond 0.3 leads to performance degradation; for instance, at $\beta = 0.5$, the MME total score drops from 643.67 to 619.50. This indicates that an excessive β risks over-correcting the output distribution, potentially compromising the model's linguistic coherence.

Analysis of Model Generalization. To validate the generalizability of PECO, we extend our evaluation across a diverse spectrum of MLLMs. As detailed in Table 5, PECO achieves consistent performance gains across all evaluated architectures. For example, on InternVL2, PECO reduces CHAIR_S and CHAIR_I by 17.6% and 4.4%, respectively, while improving POPE-F1 by 2.5%. Remarkably, even when deployed on the cutting-edge Qwen3-VL, PECO consistently curtails CHAIR_S and CHAIR_I by an additional 3.8% and 1.5%, respectively. These results demonstrate that the dual pathology of perceptual deficiency and cognitive inertia is a ubiquitous challenge inherent to MLLMs rather than a model-specific artifact. PECO serves as a robust, architecture-agnostic solution that can be seamlessly integrated into diverse MLLMs, thereby enhancing their overall reliability.

Table 5: Performance on different architectures.

Model	CHAIR_S↓	CHAIR_I↓	POPE-F1↑
Qwen2-VL	25.0	7.3	86.6
↪ **PECO**	**21.9** (-3.1)	**5.3** (-2.0)	**88.6** (+2.0)
Qwen-VL-Chat	45.6	12.5	87.0
↪ **PECO**	**43.0** (-2.6)	**10.2** (-2.3)	**88.9** (+1.9)
Qwen3-VL	19.2	5.5	89.1
↪ **PECO**	**15.4** (-3.8)	**4.0** (-1.5)	**90.9** (+1.8)
MiniGPT-4	31.8	9.9	70.3
↪ **PECO**	**29.5** (-2.3)	**9.0** (-0.9)	**71.6** (+1.3)
InstructBLIP	58.8	23.7	84.4
↪ **PECO**	**53.9** (-4.9)	**12.5** (-11.2)	**86.1** (+1.7)
Mini-Gemini	32.6	8.7	85.6
↪ **PECO**	**25.9** (-6.7)	**8.1** (-0.6)	**87.5** (+1.9)
InternVL2	45.8	12.9	86.4
↪ **PECO**	**28.2** (-17.6)	**8.5** (-4.4)	**88.9** (+2.5)

Analysis of Smoothing Kernel Size. Figure 6 presents a sensitivity analysis of the smoothing kernel size in our Spatially-Aware Importance Smoothing mechanism. We observe that a 5×5 kernel consistently achieves optimal performance. For LLaVA-1.5-7B, this configuration yields 86.9% F1 and 87.2% accuracy on POPE, outperforming both the 3×3 (86.2%/86.5%) and 7×7 (85.5%/85.9%) variants. Qwen2-VL exhibits a similar trend. These results demonstrate that a 3×3 kernel offers an inadequate receptive field to consolidate fragmented attention patterns, whereas a 7×7 kernel induces severe over-smoothing, inadvertently blurring fine-grained semantic boundaries. Notably, the performance variations across different kernel sizes are relatively modest, indicating that our spatial smoothing mechanism is effective and robust to kernel size selection.



(a) POPE-F1↑. (b) POPE-Acc.↑

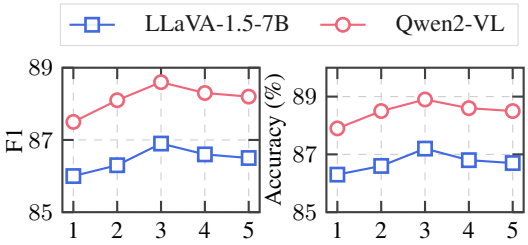
Figure 6: Sensitivity analysis of the smoothing kernel size on model performance.

Table 6: Ablation study of the perception enhancement and cognition calibration in PECO.

Method	POPE		CHAIR		MME
	F1↑	Acc.↑	$CHAIR_S$↓	$CHAIR_I$↓	Total↑
LLaVA-1.5-7B					
+ PECO	**86.9**	**87.2**	**33.3**	**8.0**	**643.67**
w/o PE	85.5	86.0	36.3	9.8	620.33
w/o CC	86.2	86.5	35.1	8.6	632.67
Qwen2-VL					
+ PECO	**88.6**	**88.9**	**21.9**	**5.3**	**686.67**
w/o PE	87.2	87.6	25.2	6.8	660.33
w/o CC	87.8	88.3	23.3	6.2	674.00

Table 7: Ablation study of the spatial and temporal modules in perception enhancement.

Method	POPE		CHAIR		MME
	F1↑	Acc.↑	$CHAIR_S$↓	$CHAIR_I$↓	Total↑
LLaVA-1.5-7B					
+ PE	**86.2**	**86.5**	**35.1**	**8.6**	**632.67**
w/o Spa.	84.8	85.0	36.6	10.2	617.33
w/o Tem.	85.4	85.5	36.0	9.1	620.67
Qwen2-VL					
+ PE	**87.8**	**88.3**	**23.3**	**6.2**	**674.00**
w/o Spa.	86.2	86.8	25.1	7.6	643.13
w/o Tem.	86.9	87.4	24.8	7.0	652.00



(a) POPE-F1↑. (b) POPE-Acc.↑

Figure 7: Sensitivity analysis of the anchor layer selection on model performance.

Analysis of Anchor Layer Selection. Figure 7 presents a sensitivity analysis of the anchor layer (L_{anchor}) within our Sensory-Deprived Contrastive Decoding framework. Empirical evaluations reveal an inverted U-shaped trajectory, with the third layer consistently securing optimal performance across both models. Specifically, on LLaVA-1.5-7B, deploying layer 3 as the anchor yields a peak 86.9% F1 and 87.2% accuracy on POPE, outperforming both shallower layers (e.g., 86.0%/86.3% at layer 1) and deeper ones (e.g., 86.5%/86.7% at layer 5). Qwen2-VL mirrors this trend, peaking at layer 3 (88.6% F1, 88.9% accuracy). This optimal positioning reflects a critical balance between visual-linguistic integration and modifiability. Layers 1-2 are too shallow to capture sufficient visual-linguistic interactions, while layers 4-5 are too deep, where the model has already committed to hallucination-prone trajectories. Layer 3 represents a sweet spot where visual semantics are sufficiently integrated yet remain modifiable, allowing our sensory deprivation mechanism to effectively isolate and penalize cognitive inertia. Notably, the performance degradation beyond layer 3 is gradual rather than abrupt, indicating that our method remains robust to anchor layer selection within a reasonable range.

5.4 Ablation Analysis

Finally, we conduct ablation experiments to verify the efficacy of each component in PECO. Table 6 summarizes the performance shifts when removing perception enhancement (PE) or cognition calibration (CC). Specifically, removing PE (w/o PE) triggers a significant performance drop on MME (e.g., -23.34 for LLaVA-1.5-7B), highlighting its role in recovering fine-grained visual details. Excluding CC (w/o CC) leads to increased generative hallucinations, with $CHAIR_S$ rising from 33.3% to 35.1% for LLaVA-1.5-7B and from 21.9% to 23.3% for Qwen2-VL, confirming its function in neutralizing linguistic cognitive inertia. PECO consistently outperforms either single-component variant, demonstrating that effective hallucination mitigation requires synergy between perception enhancement and cognition calibration. Table 7 further analyzes the respective roles of the spatial and temporal modules within the perception enhancement framework. Removing either module consistently degrades performance, indicating that a joint spatio-temporal reallocation of attention is essential to continuously anchor the generation process to the fine-grained visual evidence.

6 Conclusion

In this paper, we introduce PECO, a unified, training-free framework designed to mitigate MLLM hallucinations by synergizing visual sharpening with prior suppression. Specifically, PECO employs Spatio-Temporal Attention Rectification to progressively reclaim attention wasted on sink tokens and reallocate it to informative visual tokens along spatial and temporal dimensions, thereby reinforcing visual grounding. Concurrently, it utilizes Sensory-Deprived Contrastive Decoding to surgically isolate and penalize hallucinations driven by ingrained language priors. Extensive evaluations across diverse benchmarks and model architectures demonstrate that PECO significantly outperforms existing methods, effectively mitigating hallucinations while enhancing multimodal reasoning capabilities.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the translation of selected passages. We have not used generative AI tools to develop conceptual frameworks, formulate hypotheses or mathematical claims, design research methodology or experiments, implement the proposed method, clean or reformat datasets, or interpret experimental results; proof-related tasks, synthetic-data generation, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to edit selected portions of the manuscript for grammar, clarity, and concision. We have reviewed all AI-assisted work. Specifically, all AI-assisted text was manually checked and revised by the authors to ensure technical accuracy and consistency with the reported methods, results, and cited literature. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human participants, personally identifiable or sensitive data, or the collection or release of new datasets; therefore, institutional review board approval is not applicable. All experiments are conducted using publicly available MLLMs and established vision-language benchmarks for research purposes. The proposed method is a training-free inference-time intervention designed to mitigate hallucinations and introduces no additional privacy, security, or fairness risks beyond those inherent in the underlying models and datasets. We make no claims regarding direct deployment in safety-critical settings, where model outputs should remain subject to human verification. To the best of our knowledge, this work raises no additional concerns regarding harmful applications, legal compliance, research integrity, or conflicts of interest.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide a complete formulation of PECO and its components in Section 4. Section 5.1 reports the evaluation benchmarks, model configurations, baseline methods, implementation details, and all hyperparameter settings used in the main experiments. Sections 5.3 and 5.4 further present hyperparameter sensitivity analyses and component-wise ablations. Model-specific configurations, including the sink dimensions used for each MLLM, are documented in Appendix A, while computational overhead is detailed in Appendix B. All experiments use publicly available model checkpoints and established benchmarks, and the baseline methods are reproduced using their official implementations.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A

survey. *Findings of the association for computational linguistics: ACL 2024*, pp. 13590–13618, 2024.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Th6NyL07na.

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

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=DgH9YCsqWm.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18135–18143, 2024.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Zhiqiang Hu, Mingxing Xu, and Qixiu Cheng. Multimodal large-language model empowering next-generation autonomous driving systems. *Journal of Intelligent and Connected Vehicles*, 8(2): 9210059–1, 2025.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024a.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. Visual hallucinations of multi-modal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9614–9631, 2024b.

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

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=rsZwwjYHuD.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7uDI7w5RQA.

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

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024. URL https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

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

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd international conference on computer, artificial intelligence and control engineering*, pp. 405–409, 2024.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023a.

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

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024c.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024d.

Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

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

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

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024a.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024b.

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

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024a.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*, 2024b.

Yujun Wang, Jinhe Bi, Soren Pirk, Yunpu Ma, et al. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10306–10314, 2026.

Jinfeng Wei and Xiaofeng Zhang. Dopra: Decoding over-accumulation penalization and re-allocation in specific weighting layer. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7065–7074, 2024.

Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. Ritual: Random image transformations as a universal anti-hallucination lever in large vision language models. *arXiv preprint arXiv:2405.17821*, 2024.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S Yu. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 2247–2256. IEEE, 2023.

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

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in lvlms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3512–3534, 2025a.

Yuting Zhang, Kaishen Yuan, Hao Lu, Yutao Yue, Jintai Chen, and Kaishun Wu. Medtvt-r1: A multimodal llm empowering medical reasoning and diagnosis. *arXiv preprint arXiv:2506.18512*, 2025b.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

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

Younan Zhu, Linwei Tao, Minjing Dong, and Chang Xu. Mitigating object hallucinations in large vision-language models via attention calibration. *arXiv preprint arXiv:2502.01969*, 2025.

A Details of Sink Dimensions

In this section, we provide the specific details of the sink dimensions $\mathcal{D}_{sink}$ used in Equation 1 for various MLLMs evaluated in our experiments. Following recent findings (Sun et al., 2024a; Kang et al., 2025), LLMs exhibit massive activations in specific feature dimensions of the hidden states. These dimensions, termed *sink dimensions*, consistently possess outlier values regardless of the input tokens. Crucially, this property is inherited by MLLMs from their base LLMs and remains invariant during the multimodal alignment process (e.g., visual instruction tuning). We identify the set of sink dimensions $\mathcal{D}_{sink}$ for each model by examining the hidden states of the pre-trained LLMs. The specific indices used in our framework (as referenced in Equation 1) are listed in Table 8.

Table 8: The specific indices of massive activation dimensions ($\mathcal{D}_{sink}$) for different MLLMs. Models sharing the same base LLM architecture (e.g., Qwen/Vicuna family) share identical sink dimensions.

MLLM	**Base LLM**	$\mathcal{D}_{sink}$
LLaVA-1.5-7B	Vicuna-7B	{1415, 2533}
MiniGPT-4-7B	Vicuna-7B	{1415, 2533}
InstructBLIP-7B	Vicuna-7B	{1415, 2533}
Qwen2-VL-7B	Qwen2-7B	{458, 2570}
Qwen-VL-Chat	Qwen-7B	{458, 2570}
InternVL2-8B	Llama3-8B	{396, 1216, 2812}
Mini-Gemini-8B	Llama3-8B	{396, 1216, 2812}

B Overhead Analysis

Table 9: Efficiency comparison on NVIDIA H800 (LLaVA-1.5-7B, POPE).

Method	**Time (s)**	**Memory (MB)**	**Accuracy (%)**
Baseline	301	15673	84.0
VCD	681	16753	85.0
OPERA	1952	21943	85.2
PeCo	776	17049	87.2

To comprehensively evaluate the practical deployment viability of PeCo, we analyze its computational efficiency against standard autoregressive decoding (Baseline) and two representative state-of-the-art training-free interventions: VCD (representing the contrastive decoding paradigm) and OPERA (representing attention-penalty strategies). The evaluation is conducted on the POPE benchmark using LLaVA-1.5-7B on a single H800 GPU. As reported in Table 9, integrating PeCo naturally introduces a moderate overhead in inference time (776s) and memory consumption (17049 MB) compared to the naive baseline (301s, 15673 MB). This increase is an inherent architectural characteristic of contrastive decoding frameworks, which fundamentally necessitate dual forward passes (a standard stream and a sensory-deprived stream) to compute the penalty logits.

However, a comparative analysis confirms that this computational investment yields a disproportionately high return. When evaluated against VCD, which shares an identical dual-stream structural cost (681s, 16753 MB), PeCo incurs only a marginal additional overhead yet delivers a striking +2.2% accuracy improvement. Furthermore, when compared to OPERA—which imposes a crippling latency (1952s) and massive memory footprint (21943 MB)—PeCo operates over 2.5$\times$ faster while saving nearly 5GB of memory, all while establishing a new state-of-the-art accuracy (87.2%). Ultimately, the

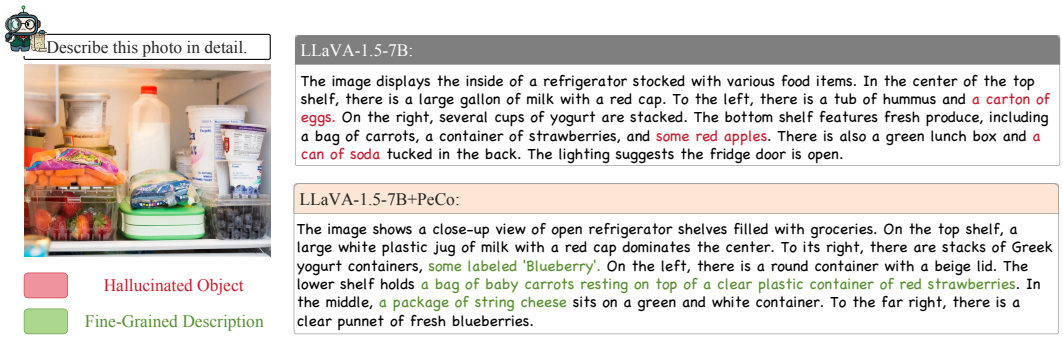


Figure 8: Qualitative comparison in a dense, cluttered environment. The baseline model suffers from situational priors, hallucinating expected items like eggs and soda. PECO thoroughly eliminates these fabrications and accurately captures complex spatial relationships (e.g., carrots resting on strawberries), demonstrating superior granular visual grounding.

substantial +3.2% performance leap over the baseline demonstrates that PECO's resource trade-off is profoundly justified, offering a highly efficient and reliable solution for MLLM deployment.

C CASE STUDY

As depicted in Figure 8, we present a dense, everyday scenario—the interior of a stocked refrigerator. This cluttered scene rigorously tests a model's susceptibility to situational language priors, as refrigerators are heavily associated with specific common grocery items in training corpora. The baseline LLaVA-1.5-7B generates a description that is linguistically plausible but factually compromised. It hallucinates several non-existent items, explicitly claiming to see "a carton of eggs," "some red apples," and "a can of soda." These severe object-level fabrications perfectly illustrate the pitfalls of cognitive inertia: when faced with a dense visual scene, the baseline abandons strict visual grounding and defaults to populating the output based on statistical "refrigerator" priors.

In sharp contrast, equipping the model with PECO completely eradicates these ungrounded fabrications. PECO strictly adheres to the visual evidence, successfully avoiding the phantom eggs, apples, and soda. Furthermore, it demonstrates dramatically enhanced fine-grained perception and spatial reasoning capabilities. For instance, it accurately reads the text on the packaging (noting the yogurt is "labeled Blueberry") and precisely articulates complex object interactions, such as "a bag of baby carrots resting on top of a clear plastic container of red strawberries." This example validates that PECO successfully neutralizes ingrained situational priors, ensuring that the model's generation remains anchored to granular, real-world visual evidence even in highly complex environments.

D MORE DISCUSSION

Role of Sink Tokens in Sensory Deprivation. In our Sensory-Deprived Contrastive Decoding framework, we preserve visual sink tokens rather than removing all visual information when constructing the blind stream (V^-). This design choice may seem counterintuitive at first glance, as one might expect complete visual removal to achieve true sensory deprivation. However, we find that retaining sink tokens is crucial for structural stability rather than semantic leakage. Conceptually, sink tokens are characterized as semantic voids; they absorb massive attention solely due to the softmax normalization mechanism, encoding little to no content regarding the image's foreground objects. Therefore, their presence does not reintroduce the visual semantics we aim to deprive.

Moreover, functionally, these tokens serve as computational anchors that regulate the dynamic range of attention scores in pre-trained MLLMs. Completely stripping them would induce a catastrophic distribution shift, causing the blind stream (V^-) to degenerate into incoherent gibberish rather than generating the fluent, logic-driven hallucinations we aim to capture and penalize. Preserving sink tokens thus ensures that the contrastive stream represents a functioning language model operating without visual evidence, rather than a "broken" model.

Training-Free vs. Training-Based Approaches. PECO adopts a training-free intervention strategy, in contrast to fine-tuning methods (e.g., RLHF, DPO) that directly optimize model behavior through parameter updates. This design choice offers distinct advantages that complement rather than replace training-based methods. First, PECO provides immediate deployment capability without requiring access to proprietary training infrastructure or curated preference datasets, which are often expensive and domain-specific. Second, our framework preserves the model's original parametric knowledge, avoiding catastrophic forgetting that can occur during fine-tuning. Third, PECO can be seamlessly combined with training-based methods: one could first fine-tune a model for domain-specific alignment, then apply PECO during inference to further suppress residual hallucinations. Moreover, training-based methods struggle with out-of-distribution scenarios where the training data coverage is limited, whereas PECO's attention-level interventions adapt dynamically to each input.

Justification for Attention Reclamation Strategy. In our attention reclamation mechanism, we redistribute attention from sink tokens to non-sink tokens rather than simply removing sink tokens entirely. This design choice is grounded in attention budget conservation: in Transformer architectures, attention weights sum to 1 due to softmax normalization. Simply zeroing out sink tokens would cause the remaining attention mass to redistribute proportionally, potentially amplifying existing biases. By explicitly reclaiming and reallocating the attention budget using spatially-smoothed importance maps, we ensure that the redistributed mass flows to semantically meaningful regions rather than being absorbed by the next-most-attended (potentially noisy) tokens. This ensures that isolated high-attention tokens do not monopolize the reclaimed budget, promoting coherent semantic clusters that align with visual object boundaries.

E LIMITATIONS

While PECO demonstrates robust performance across various benchmarks, there are two aspects worth noting for future exploration. First, PECO operates as a white-box intervention strategy, necessitating access to the model's internal attention distributions and output logits. Consequently, our framework is designed for open-source models and cannot be directly applied to closed-source commercial MLLMs (e.g., GPT-4V (Achiam et al., 2023)) that are accessible solely via black-box APIs. Second, our current implementation specifically targets the attention sink phenomenon observed in standard Transformer-based architectures. As the field evolves towards diverse architectural paradigms, such as State Space Models (Liu et al., 2024d) or linear attention mechanisms, the applicability and adaptation of our attention rectification strategy to these emerging non-Transformer architectures remain an intriguing direction for future research.

F BROADER IMPACTS

The development of PECO presents significant positive societal impacts by substantially enhancing the reliability and safety of MLLMs. By effectively mitigating hallucinations, our training-free framework facilitates the responsible deployment of MLLMs in high-stakes domains that demand rigorous factual grounding, such as medical image analysis, autonomous navigation, and visually-assisted robotics. Furthermore, as a plug-and-play decoding intervention, PECO promotes Green AI. It democratizes access to robust multimodal reasoning, allowing practitioners to explicitly improve existing models without incurring the massive environmental footprint and computational costs associated with full-scale retraining.

Despite these benefits, PECO does not eliminate all risks associated with MLLMs. As model outputs become more visually grounded, users may place greater trust in the system, so human review remains important in sensitive or high-stakes applications. In addition, since PECO operates at the decoding and attention-modulation levels, it does not directly address biases that may already exist in the underlying pre-trained models. We therefore view PECO as a complementary reliability improvement, and future work may combine hallucination mitigation with dedicated bias-reduction techniques.