

Disentangling Semantic Attention from Structural Bias in the Attention Manifold

Pengkun Jiao
Fudan University
Shanghai, China
pkjiao23@m.fudan.edu.cn

Jingjing Chen*
Fudan University
Shanghai, China
chenjingjing@fudan.edu.cn

Bin Zhu
Singapore Management University
Singapore
binzhu@smu.edu.sg

Yu-Gang Jiang
Fudan University
Shanghai, China
ygj@fudan.edu.cn

Abstract

The empirical success of attention mechanism in Multimodal Large Language Models (MLLMs) often obscures its inherent, subtle flaws. Specifically, MLLMs consistently exhibit disproportionate attention toward certain semantically uninformative visual tokens, a phenomenon termed “register” or “Visual Attention Sinks.” While existing inference intervention methods attempt to identify these sink tokens and redistribute their attention weights, such approaches typically treat these tokens in isolation and suffer from computational inefficiency. Instead, we reframe this phenomenon as a generalized textual bias exerted over visual features that extends beyond isolated sink tokens. From this perspective, a pervasive structural bias leads to the dilution of the semantic visual signal, precipitating multimodal hallucinations as the model prioritizes linguistic priors over valid visual evidence. To address this limitation, we introduce Saliency-guided Purification and Adaptive Redistribution (SPAR), a training-free, plug-and-play intervention. SPAR mitigates this generalized textual bias by purifying structural noise and subsequently redistributing the reclaimed attention budget to the most informative visual regions. Comprehensive evaluations across a diverse spectrum of hallucination benchmarks demonstrate that SPAR effectively restores authentic visual grounding with negligible computational overhead.  Project Page

CCS Concepts

• **Computing methodologies** → **Computer vision**; **Natural language processing**.

Keywords

multimodal large language models, attention, structural bias, hallucination, posterior collapse, linguistic prior

ACM Reference Format:

Pengkun Jiao, Bin Zhu, Jingjing Chen, and Yu-Gang Jiang. 2026. Disentangling Semantic Attention from Structural Bias in the Attention Manifold. In

*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License. MM '26, Rio de Janeiro, Brazil.

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2213-4/2026/11

<https://doi.org/10.1145/3767308.3836288>

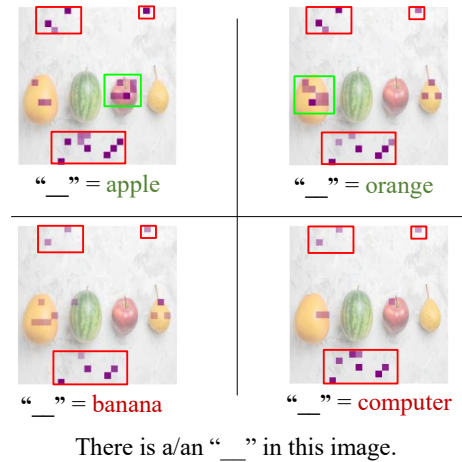


Figure 1: Visualization of structural attention noise. High attention responses consistently concentrate on fixed image regions across different textual queries, indicating a query-invariant bias.

Proceedings of the 35th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 21 pages. <https://doi.org/10.1145/3767308.3836288>

1 Introduction

The rapid advancement and successful deployment of Multimodal Large Language Models (MLLMs) across diverse real-world applications have highlighted their transformative potential [4, 13, 15, 19, 31]. However, their overall reliability is persistently compromised by hallucinations—where models produce responses inconsistent with the visual evidence—significantly hindering their safe and dependable adoption [2, 20].

Traditionally, literature has focused on *Spontaneous Hallucinations* [20, 25], which arise from internal model noise or an over-reliance on statistical priors. Yet, as MLLMs are increasingly deployed in complex, interactive environments, a more severe vulnerability has emerged: *Induced Hallucinations*, or *Gaslighting* [36]. In these scenarios, misleading external cues, e.g., deceptive user prompts, compel the model to abandon its visual grounding entirely, prioritizing linguistic compliance over visual fidelity [36].

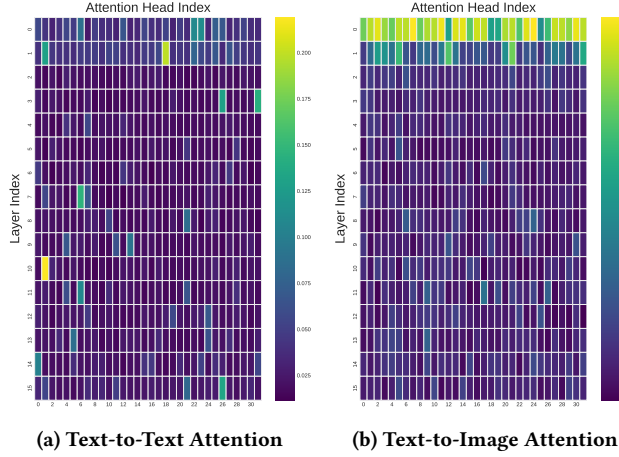


Figure 2: Mean attention maps of (a) text-to-text and (b) text-to-image interactions in LLaVA-NeXT. Structural bias is substantially stronger in text-to-image attention, especially in the first two layers.

An investigation into the internal dynamics of MLLMs reveals a structural flaw: MLLMs frequently exhibit persistent, query-agnostic attention artifacts. As illustrated in Figure 1, rather than focusing on task-relevant objects, the model’s attention is often misdirected toward fixed spatial regions or semantically uninformative tokens. Recent studies have characterized this behavior as *registers* [7] or *visual attention sinks* [16], treating these anomalous attention patterns as isolated receptacles to absorb redundant attention. However, we argue that these identified sinks are merely localized symptoms of a broader phenomenon: a generalized text-to-visual attention bias.

We posit that the attention manifolds in MLLMs exhibit a static, latent *Structural Bias*. Mathematically, this stems from the softmax normalization constraint, which forces the model to distribute residual probability mass across non-informative regions to satisfy the unit-sum requirement. By analogy to Polya’s Urn dynamics [24], we hypothesize that the model reinforces subtle initial asymmetries—such as those inherent in rotary positional encodings [26]—until they crystallize into fixed positional biases. As shown in Figure 2, text tokens maintain a high average attention toward fixed image positions across initial layers, where visual and linguistic modalities have not yet deeply interacted, allowing these asymmetric structural priors to easily dominate. Consequently, the *Semantic Attention* signal is eclipsed by this high-magnitude structural bias, rendering the model highly susceptible to hallucination.

Current inference-time interventions exhibit critical limitations. Contrastive decoding methods [5, 18] typically require multiple forward passes per generated token, fundamentally slowing down generation. More recent attention-calibration strategies [14, 16] successfully mitigate hallucinations by reallocating attention away from isolate non-essential “sink” tokens; however, they also introduce substantial inference latency, limiting their practicality for latency-sensitive applications. Furthermore, localized interventions that merely mask or sequester discrete sink tokens fail to address the underlying generalized structural biases.

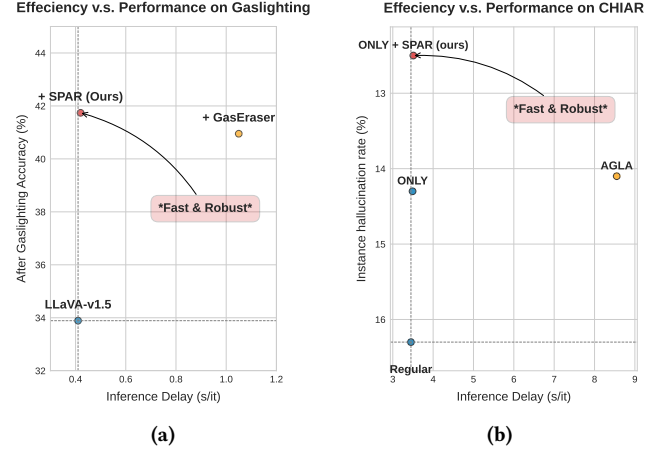


Figure 3: SPAR improves robustness to both (a) spontaneous hallucinations and (b) induced hallucinations while incurring minimal inference overhead.

To address these issues, we propose a Signal-Noise Decomposition perspective, asserting that robust grounding requires decoupling rigid structural bias from attention manifold. Building on this insight, we introduce **S**aliency-guided **P**urification and **A**daptive **R**edistribution (**SPAR**). Unlike prior methods that target isolated sinks, SPAR adopts a holistic approach by treating structural bias as a generalized noise layer superimposed on the attention manifold. As shown in Figure 3, evaluations confirm that SPAR provides a highly efficient, training-free solution that substantially enhances MLLM robustness against both spontaneous hallucinations and induced gaslighting, bridging the gap between cross-modal fidelity and low-latency inference demands.

Our main contributions can be summarized as:

- ❶ We formalize the Semantic-Structural Decoupling hypothesis, demonstrating that generalized structural biases act as static noise that dilutes semantic signals.
- ❷ We propose SPAR, a training-free, plug-and-play mechanism that leverages visual saliency to decouple structural bias and restore visual SNR.
- ❸ We demonstrate SPAR’s effectiveness across diverse benchmarks, achieving competitive performance in mitigating both spontaneous and induced hallucinations with negligible latency.

2 Related Works

Attention Sink. describes the tendency of LLMs to allocate disproportionately high attention to a small subset of tokens—typically the initial few—regardless of their semantic contribution [32]. Xiao et al. [32] demonstrated that this behavior is consistent and gives rise to a strong attentional bias toward these early tokens. Subsequent research has sought to uncover the underlying causes of this phenomenon. Cancedda et al.[3] proposed that the attention sink is often concentrated in the very first token, attributing this bias to the large norm of its hidden state. In contrast, other studies have observed that attention sinks can manifest in various

semantically weak tokens without a fixed position in the input sequence [27, 33], complicating the attention distribution. A similar phenomenon, referred to as *Visual Attention Sink*, has also been observed in MLLMs. Timothée *et al.* [7] showed that high-norm tokens frequently emerge during inference in vision transformers, primarily from low-information background regions of images. Seil *et al.* [16] further investigated this phenomenon in MLLMs and proposed an attention reallocation strategy to mitigate it.

Hallucination in MLLMs. MLLMs may generate responses that hallucinate content not grounded in the visual input, resulting in misinformation and degraded performance [18, 20]. Early attempts to address this issue have relied on additional robust instruction tuning using specially curated datasets [11, 28]. Although these approaches can be effective, they demand substantial training resources, making them impractical for most users. Consequently, recent efforts have shifted toward training-free inference-time interventions. One dominant category is contrastive decoding-based methods, which mitigate hallucinations by contrasting logits from paired outputs [1, 5, 8, 18]. While promising, these methods typically require multiple queries, slowing down generation, though a very recent work [30] attempts to optimize this via a single-query, one-layer intervention. Parallel to logit-level manipulation, another emerging line of research focuses on refining internal attention mechanisms and visual grounding. For example, PAI [22] directly enhances visual fidelity by compelling the model to pay more attention to image tokens during decoding without any extra training. To further interpret these internal dynamics, HGAI [12] utilizes an attention lens to identify and rectify object hallucinations originating in the middle layers of MLLMs. From a causal perspective, CasualMM [35] mitigates hallucinations by deciphering attention causality to suppress modality priors that often lead to erroneous predictions. Moreover, MemVR [37] introduces a memory-space visual retracing mechanism, enabling the model to “look twice” at the visual input to verify its responses. More recently, attention has turned toward induced hallucinations triggered by misleading prompts. GaslightBench [36] established a benchmark for evaluating model resilience against linguistic pressure (i.e., gaslighting). To address this, GasEraser [14] was proposed to mitigate these effects by relocating anomalous attention patterns specifically triggered under such pressure. GasVideo [29] further extends negation-based gaslighting to video LLMs.

3 Semantic-Structural Decoupling

In this section, we formalize the mechanism of visual hallucination through the lens of Information Theory and Bayesian Inference. We propose the **Semantic-Structural Decoupling (SSD)** perspective, which posits that the attention manifold is a superposition of latent semantic attentions and rigid structural bias. We demonstrate that hallucinations become inevitable when the rigid structural bias overshadows the semantic signal.

3.1 Preliminaries: MLLMs and Softmax Attention

Multimodal Large Language Models (MLLMs), such as LLaVA [21], typically integrate a visual encoder f_ϕ , a modality projector g_θ , and a Transformer-based LLM h_ψ .

Given an image I , the visual features are mapped into the word embedding space as tokens $X_v = \{x_{v,1}, \dots, x_{v,n}\}$, which are concatenated with textual tokens X_t to form the input sequence $X = [X_v; X_t]$. These tokens are processed via softmax attention. For a given head, the attention matrix $A = [\alpha_{ij}]$ is computed as:

$$A = \text{Softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right), \text{ where } \alpha_{ij} = \frac{\exp(q_i k_j^T / \sqrt{d_k})}{\sum_l \exp(q_i k_l^T / \sqrt{d_k})}. \quad (1)$$

The softmax operation imposes a unit-sum constraint, i.e., $\sum_j \alpha_{ij} = 1$, which forces attention mass to be allocated even when the query has weak semantic correspondence with the available visual tokens. Furthermore, the Exponential Amplification property [38] of Softmax exacerbates this issue by magnifying even minor variations in key norms (e.g., positional encoding asymmetry). Ultimately, this causes high-variance visual tokens to disproportionately dominate the attention distribution, emerging as attention ‘sinks’ [16].

3.2 Manifold Decomposition: The Signal-Noise Hypothesis

The unit-sum constraint in standard Softmax forces the model to distribute probability mass even in the absence of correlation. Consequently, the resulting attention map may conflate true semantic importance with normalization artifacts. To formalize this, we model the observed attention matrix $A_{obs} \in \mathbb{R}^{|T| \times |V|}$ as a superposition of a latent semantic signal and a structural bias component.

Definition 1 (Structural Bias). We define structural bias, denoted by A_{struct} , as the query-invariant component of the attention manifold. It captures the inherent spatial preference of an attention head independent of the specific text query:

$$A_{struct}(V) \triangleq \mathbb{E}_{q \sim P(T)} [A_{obs}(q, V)] \approx \frac{1}{N} \sum_{i=1}^N A_{obs}(i, \cdot) \quad (2)$$

Definition 2 (Semantic Attention). We define semantic attention, denoted by A_{sem}^* , as the latent alignment distribution that reflects the true semantic correspondence between textual queries and visual features, disentangled from structural bias.

Based on the above definitions, we posit an additive noise model for the observed attention:

$$A_{obs}(T, V) = \underbrace{A_{sem}^*(T, V)}_{\text{True Signal}} + \underbrace{\lambda \cdot A_{struct}(V)}_{\text{Structural Bias}} + \zeta, \quad (3)$$

where λ is a modulation coefficient capturing the varying susceptibility to structural bias across different attention heads. The term ζ represents the *Stochastic Approximation Error*, encompassing approximation errors and high-entropy noise that, while not part of the query-invariant structural bias, contributes no discriminative semantic information to the cross-modal alignment. Further empirical analysis of structural bias is detailed in Appendix B.

3.3 Inference as Iterative Bayesian Update

We interpret the autoregressive generation process of MLLMs through the lens of iterative Bayesian updating. For a candidate next-token $y \in \mathcal{Y}$, the posterior probability is given by:

$$P(y|V, T) = \frac{P(V|y, T) \cdot P(y|T)}{\sum_{y' \in \mathcal{Y}} P(V|y', T) \cdot P(y'|T)}, \quad (4)$$

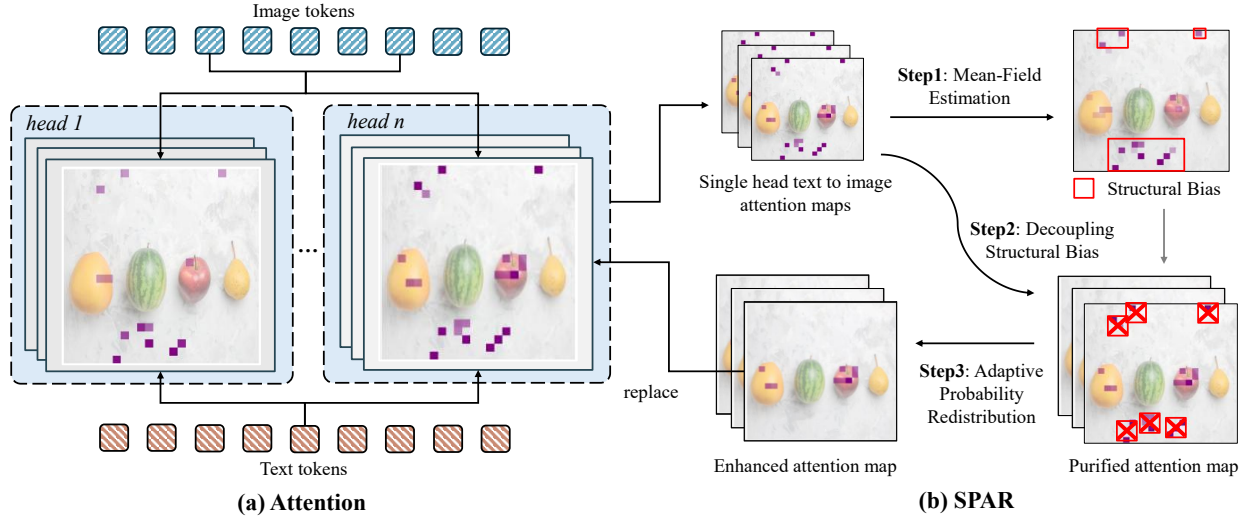


Figure 4: Illustration of SPAR Pipeline. (a) Text-to-image cross-attention maps are extracted from all attention heads. (b) A fixed structural bias is derived from high-frequency patterns within the attention maps. This bias is then subtracted to purify the attention distribution, and the residual probability mass is adaptively redistributed to the remaining regions to reinforce visual attention.

where:

- $P(y|T)$ is the **Linguistic Prior**, derived from the LLM’s pre-trained corpus and the current textual context (including system instructions and user prompts).
- $P(V|y, T)$ is the **Visual Likelihood**, representing the probability that the visual evidence V supports the hypothesis of token y .

3.4 Posterior Collapse

To quantify the relative visual support across the vocabulary $\mathcal{Y}$, we define the normalized visual likelihood

$$L(y) = \frac{P(V|y, T)}{\sum_{y' \in \mathcal{Y}} P(V|y', T)}. \quad (5)$$

Let c_y denote the visual context vector aggregated for evaluating candidate token y . When the attention manifold is dominated by structural bias $\mathbf{A}_{struct}$ and residual noise ζ , the resulting visual context becomes increasingly insensitive to token-specific semantics. In that regime, the visual likelihood loses discriminability across competing candidates. To characterize this degradation, we define the visual signal-to-noise ratio (SNR) as

$$\text{SNR} \triangleq \frac{\|\mathbf{A}_{sem}^*\|_F^2}{\|\lambda \mathbf{A}_{struct} + \zeta\|_F^2}, \quad (6)$$

where $\|\cdot\|_F$ denotes the Frobenius norm, representing the total energy concentrated in each component of the attention manifold.

Corollary 1 (Posterior Collapse Toward the Linguistic Prior). *Assume that, as $\text{SNR} \rightarrow 0$, the normalized visual likelihood $L(y)$ becomes approximately non-discriminative across candidate tokens, i.e., there exists a continuous function $\varepsilon(\text{SNR}) \geq 0$ with $\lim_{\text{SNR} \rightarrow 0} \varepsilon(\text{SNR}) = 0$ such that*

$$\sup_{y, y' \in \mathcal{Y}} |L(y) - L(y')| \leq \varepsilon(\text{SNR}). \quad (7)$$

Then the posterior distribution approaches the linguistic prior, and

$$D_{KL}(P(y|V, T) \parallel P(y|T)) \leq f(\text{SNR}), \quad (8)$$

where $f(\cdot)$ is non-negative, continuous, and satisfies $\lim_{\text{SNR} \rightarrow 0} f(\text{SNR}) = 0$.

The proof sketch of Corollary 1 is provided in Appendix A. This corollary provides a unified explanation for both spontaneous and induced hallucinations:

- **Spontaneous Hallucination:** This phenomenon occurs when the visual likelihood $P(V|y, T)$ is attenuated by internal structural bias. As the SNR decreases, the visual evidence fails to produce a discriminative distribution, causing the posterior to collapse toward the *Linguistic Prior* $P(y|T)$. In this state, the generative process is governed by the statistical density of the pre-training corpus rather than the provided visual features.
- **Induced Hallucination:** This occurs when a deceptive textual context T_{adv} biases the linguistic prior $P(y|T_{adv})$ toward a specific candidate subspace. Under structural saturation (low SNR), the low-magnitude visual likelihood is insufficient to counteract the high-probability mass of the biased prior.

4 Methodology

A straightforward strategy for mitigating the degradation of visual discriminability is to isolate and remove the deterministic structural bias inherent in the attention manifold. To this end, we propose **SPAR** (Saliency-guided Purification and Adaptive Redistribution). By purifying the attention manifold, **SPAR** effectively increases the SNR of the visual likelihood, thereby preventing the posterior distribution from collapsing toward the linguistic prior.

Given the cross-modal attention mechanism in an MLLM layer, let $\mathbf{T}$ and $\mathbf{V}$ denote the sets of text and visual tokens, respectively. The attention matrix $\mathbf{A} \in \mathbb{R}^{|\mathbf{T}| \times |\mathbf{V}|}$ represents the correlation between text queries and visual keys. We observe that in early layers,



Figure 5: Qualitative visualization of attention purification via SPAR on LLaVA-v1.5. (Left/Middle) Average text-to-image attention maps reveal that the baseline model suffers from persistent, semantically irrelevant biases, which are effectively rectified by SPAR. (Right) Token-level analysis demonstrates that SPAR suppresses structural bias and sharpens the attention focus on corresponding visual entities.

A often exhibits a “spurious template” pattern, where text tokens uniformly attend to high-variance visual regions regardless of actual semantic alignment.

A straightforward intervention would involve the global suppression of high-frequency attention; however, such a naive approach introduces two critical risks. First, it may lead to semantic erasure, inadvertently degrading visual signals that must remain concentrated to support multiple textual tokens (e.g., “a red silk skirt”). Second, it risks noise amplification by indiscriminately elevating low-frequency noise attention (i.e., ζ). To address these challenges, we propose a two-stage framework: **Saliency-Guided Structural Bias Purification** (Sec. 4.1) and **Adaptive Probability Redistribution** (Sec. 4.2).

4.1 Saliency-Guided Structural Bias Purification

Our method relies on the hypothesis that visual tokens with extremely high variance in the feature space act as attention sinks [16, 27]. We quantify this using a saliency map derived from the visual hidden states $\mathbf{h} \in \mathbb{R}^{|V| \times D}$.

For each visual token j , we calculate the feature variance σ_j^2 . To ensure numerical stability and comparability across batches, we normalize this variance:

$$\sigma_j^2 = \text{Var}(\mathbf{h}_j), \quad \hat{\sigma}_j = \frac{\sigma_j^2 - \mu_\sigma}{\delta_\sigma + \epsilon}, \quad (9)$$

where μ_σ and δ_σ are the mean and standard deviation of the variances across the visual sequence, and ϵ is a small constant (e.g., $1e^{-6}$). The saliency score $S_j \in (0, 1)$ is then obtained via a sigmoid activation:

$$S_j = \frac{1}{1 + e^{-\hat{\sigma}_j}}. \quad (10)$$

Then we estimate Structural Bias Template $\bar{\mathbf{A}}$ by marginalizing over the textual dimension of $\mathbf{A}$:

$$\bar{\mathbf{A}}_j = \mathbb{E}_{i \sim T} [\mathbf{A}_{i,j}] \approx \frac{1}{|T|} \sum_{i=1}^{|T|} \mathbf{A}_{i,j}. \quad (11)$$

This template captures the persistent spatial preferences of the attention head, where high values in $\bar{\mathbf{A}}_j$ indicate visual regions acting as “attractors” regardless of query semantics. To purify the attention $\mathbf{A}$, we subtract this structural bias. However, to avoid harming genuine visual perception (e.g., background context), we modulate the subtraction using the visual feature saliency S :

$$\mathbf{A}_{i,j}^{pur} = \text{ReLU}(\mathbf{A}_{i,j} - \lambda \cdot (1 + \beta S_j) \cdot \bar{\mathbf{A}}_j), \quad (12)$$

where λ and β are hyperparameters. This operation effectively decouples the structural bias from the attention map, exposing the underlying semantic signals.

4.2 Adaptive Probability Redistribution

The purification process inevitably reduces the total probability mass (energy) of the attention distribution. Simply re-normalizing the distribution may amplify residual noise; instead, SPAR employs a confidence-aware redistribution strategy. We first evaluate the *confidence* of the purified attention based on its sparsity. For a given text token i , we calculate the sparsity ratio ρ_i and the resulting confidence score γ_i :

$$\rho_i = \frac{\max_j(\mathbf{A}_{i,j}^{pur})}{\sum_j \mathbf{A}_{i,j}^{pur} + \epsilon}, \quad \gamma_i = \text{Sigmoid}(k \cdot (\rho_i - \tau_{sparsity})), \quad (13)$$

where $\tau_{sparsity}$ is the sparsity threshold, a high ρ_i indicates that the model has successfully localized a specific region after purification. We then calculate the removed probability budget $\Delta_i = \sum_j \mathbf{A}_{i,j} - \sum_j \mathbf{A}_{i,j}^{pur}$. We restore this budget proportional to the confidence γ_i , ensuring that only confident semantic signals are enhanced:

$$r_i = \frac{\sum_j \mathbf{A}_{i,j}^{pur} + \Delta_i \cdot \gamma_i}{\sum_j \mathbf{A}_{i,j}^{pur} + \epsilon}. \quad (14)$$

The final attention map is computed as $\mathbf{A}_{i,j} = \mathbf{A}_{i,j}^{pur} \cdot r_i$. This step effectively strengthens weak-but-correct semantic alignments while keeping ambiguous noise suppressed.

Table 1: Performance evaluation on the object hallucination benchmark POPE and image captioning hallucination benchmark CHAIR using LLaVA-v1.5. For caption generation in CHAIR, max generation length is set to 128.

Method	Publish Venue / Year	POPE				CHAIR	
		Precision $\uparrow$	Recall $\uparrow$	F1 Score $\uparrow$	Accuracy $\uparrow$	CHAIR _S $\downarrow$	CHAIR _I $\downarrow$
Regular [21]	CVPR 2024	71.76	90.42	79.72	76.69	55.0	16.3
+ SPAR (ours)	-	85.77	85.72	85.42	85.20	49.0	14.0
AGLA[1]	CVPR 2025	74.55	93.82	82.61	79.66	47.6	14.1
+ SPAR (ours)	-	75.47	94.13	83.70	80.70	33.4	11.9
ONLY [30]	ICCV 2025	75.00	92.87	82.60	79.87	49.8	14.3
+ SPAR (ours)	-	76.20	93.27	83.84	81.16	31.4	12.5
VAR [16]	ICLR 2025	84.14	85.69	84.61	84.25	43.2	13.8
+ SPAR (ours)	-	84.11	86.71	85.59	85.19	40.4	12.4

Table 2: Evaluation of induced hallucination on GaslightingBench. “Before-Negation” reports accuracy on the original question, and “After-Negation” reports accuracy after the negation-based gaslighting prompt is applied.

Method	Budget Source	Inference Delay	Before-Negation Acc(%)	After-Negation Acc(%)
LLaVA-v1.5	None	0.41s/it	63.25	24.71
+ Preemptive Prompt Hardening	None	0.41s/it	63.25	33.89
+ GasEraser [14]	sink token	1.05s/it	61.07	40.95
+ SPAR (ours)	structural bias	0.42s/it	63.71	41.74
LLaVA-NeXT	None	0.81s/it	65.89	19.81
+ Preemptive Prompt Hardening	None	0.81s/it	65.89	24.02
+ GasEraser [14]	sink token	1.68s/it	65.04	30.58
+ SPAR (ours)	structural bias	0.82s/it	65.30	34.04

5 Experiments

We evaluate our method on two distinct categories of visual inconsistency of MLLMs: **Spontaneous Hallucination** and **Induced Hallucination**.

Benchmarks. To holistically evaluate *spontaneous hallucinations*, we employ three complementary metrics. We utilize POPE [20] and the hallucination subset of MME [9] for discriminative assessment, conducting binary probing on object existence and attribute fidelity. For generative contexts, we apply the CHAIR [25] metric to quantify the consistency of objects referenced in image captions.

To assess *induced hallucinations*, we employ gaslighting attacks as the primary provocation mechanism. Our core evaluation utilizes GaslightingBench [36], a specialized benchmark designed to measure the resilience of multimodal models against deceptive prompts. To further validate the generalizability of our approach across diverse domains, we extend our evaluation to broader benchmarks, including MMMU [34], AI2Diagram [17], and MMBench [23]. More details about datasets and metrics are provided in Appendix C.

Baselines. We evaluate our method on two prominent open-source MLLMs: **LLaVA-1.5-7B** [21], and **LLaVA-NeXT** [19].

For the *spontaneous hallucinations* evaluation, we benchmark our method against several state-of-the-art approaches, including **OPERA** [10], **DOLA** [6], **VCD** [18], **AGLA** [1], **ONLY** [30], and

VAR [16]. We also employ multinomial sampling as the default decoding strategy (denoted as **Regular**). For the *induced hallucinations* (gaslighting) task, we compare against **GasEraser** [14], designed to mitigate the influence of misleading user prompts and ensure adherence to visual evidence. Further details regarding baselines and evaluation metrics are provided in Appendix F.

Implementation and Hyperparameters. All experiments were conducted on NVIDIA RTX A6000 GPUs. For both LLaVA-v1.5 and LLaVA-NeXT, we designate the initial two Transformer layers for **SPAR** injection. We configure the hyperparameters as $\lambda = 1.0$, $\beta = 1.0$, and $\tau_{\text{sparsity}} = 0.1$. A comprehensive sensitivity analysis regarding these parameters is provided in Appendix G.

5.1 Evaluation on Spontaneous Hallucination

Table 1 reports the results on the POPE and CHAIR benchmarks using LLaVA-v1.5. Overall, SPAR consistently reduces hallucinations across different decoding strategies, demonstrating its effectiveness as a general inference-time intervention.

Improvement over standard decoding. Compared with the Regular decoding baseline, SPAR brings substantial gains on both discriminative and generative hallucination benchmarks. On POPE, SPAR improves Accuracy from 76.69 to 85.20 and F1 score from 79.72 to

Table 3: Additional evaluation on induced hallucination. “Before” and “After” denote the model accuracy before and after the gaslighting prompt, respectively.

Dataset	Metric	LLaVA-v1.5	+ GasEraser	+ SPAR (Ours)
MMMU	Before	37.56	33.87	36.47
	After	22.14	25.86	27.96
AI2Diagram	Before	49.66	42.79	49.08
	After	29.48	32.39	36.77
MMBench	Before	72.07	68.02	72.19
	After	26.80	41.77	46.05
Average	Before	53.09	48.23	52.58
	After	26.14	33.34	36.93

85.42, corresponding to gains of 8.51 and 5.70 percentage points, respectively. On CHAIR, SPAR reduces CHAIR_s from 55.0 to 49.0 and CHAIR_l from 16.3 to 14.0. These results indicate that suppressing structural bias in text-to-image attention can effectively improve visual grounding and reduce the generation of unsupported objects. Detailed results on the three POPE subsets are provided in Appendix I.

Compatibility with existing methods. Another important advantage of SPAR is its compatibility with existing inference-time hallucination mitigation methods. We combine SPAR with several strong baselines, including AGLA [1], ONLY [30], and VAR [16]. As shown in Table 1, SPAR consistently improves these methods across both POPE and CHAIR. In particular, although VAR already achieves strong performance, **VAR + SPAR** further improves the POPE F1 score to 85.59 and reduces CHAIR_l to 12.4, yielding the best overall results in the table. This suggests that SPAR captures structural attention artifacts that are not fully addressed by prior interventions, making it a complementary module for existing decoding-based approaches.

5.2 Evaluation on Induced Hallucination

Gaslighting attacks pose a severe challenge to MLLMs by leveraging linguistic pressure to induce visual hallucinations. The evaluation follows a two-stage protocol: the model first answers a neutral baseline question, and is subsequently subjected to a negation-based gaslighting prompt to assess its resilience against adversarial suggestion. Further details regarding the gaslighting task are provided in Appendix E.

Table 2 summarizes the results, underscoring the superior efficacy of our proposed SPAR method compared to the state-of-the-art GasEraser across multiple foundational MLLMs. SPAR consistently achieves a distinct advantage in post-negation accuracy across all tested models: reaching 41.74% on LLaVA-v1.5 (vs. 40.95% for GasEraser), and 34.04% on LLaVA-NeXT (vs. 30.58%). Crucially, this enhancement in adversarial robustness does not occur at the expense of the models’ intrinsic capabilities; the pre-negation accuracy is consistently preserved or even slightly improved (e.g., 63.71% vs. 63.25% for LLaVA-v1.5). Extended evaluations regarding the gaslighting task across the MMMU, AI2Diagram, and MMBench benchmarks are provided in Appendix H.

Table 4: Ablation on the marginalization source for structural bias estimation. We report the before- and after-negation accuracy of LLaVA-v1.5 under different choices of image- and text-side marginalization.

Image Marg.	Text Marg.	Before Negation	After Negation
×	×	63.25	33.89
✓	×	63.50	34.28
×	✓	63.71	41.74
✓	✓	64.10	41.18

Table 5: Ablation study of the key components of SPAR on GaslightingBench.

Method	Before Negation	After Negation
SPAR (Full)	63.71	41.74
w/o Saliency Guidance	62.10	35.20
w/o Probability Redistribution	60.78	33.91
w/o Sparsity Guidance	57.86	36.13

We further extended our evaluation to three additional benchmarks: MMMU [34], AI2Diagram [17], and MMBench [23]. As presented in Table 9, this broader assessment confirms that the performance advantages of SPAR are consistent across varying task domains. On average, SPAR achieves a post-attack accuracy of 36.93%, representing a significant improvement over both the LLaVA-v1.5 baseline (26.14%) and GasEraser (33.34%). This performance margin is maintained across every individual dataset, affirming that SPAR provides robust protection against induced hallucinations regardless of the underlying data distribution.

5.3 Ablation Study

Structural bias primarily resides in text-to-image attention. Table 4 evaluates different marginalization sources for structural bias estimation. Using only image-side marginalization brings limited improvement, whereas text-side marginalization yields a much larger gain and performs comparably to joint marginalization. This is consistent with our formulation of structural bias as a query-invariant text-to-image attention pattern. Since hallucinations mainly stem from cross-modal attention misalignment, marginalizing over text tokens provides a more faithful estimate of the bias that SPAR is designed to remove.

Ablation Study on Key Components. Table 5 presents the ablation analysis of SPAR on GaslightingBench, confirming the necessity of each constituent module. Removing Sparsity Guidance incurs the most substantial performance degradation in the Before Negation task, with accuracy decreasing by 5.85 points. This decline highlights the component’s critical function in suppressing background noise and maintaining object-centric focus. Additionally, excluding Probability Redistribution leads to the most severe drop in After Negation performance (falling to 33.91), demonstrating its essential role in preserving semantic consistency under logical inversions. Collectively, optimal performance requires the synergy of all components.

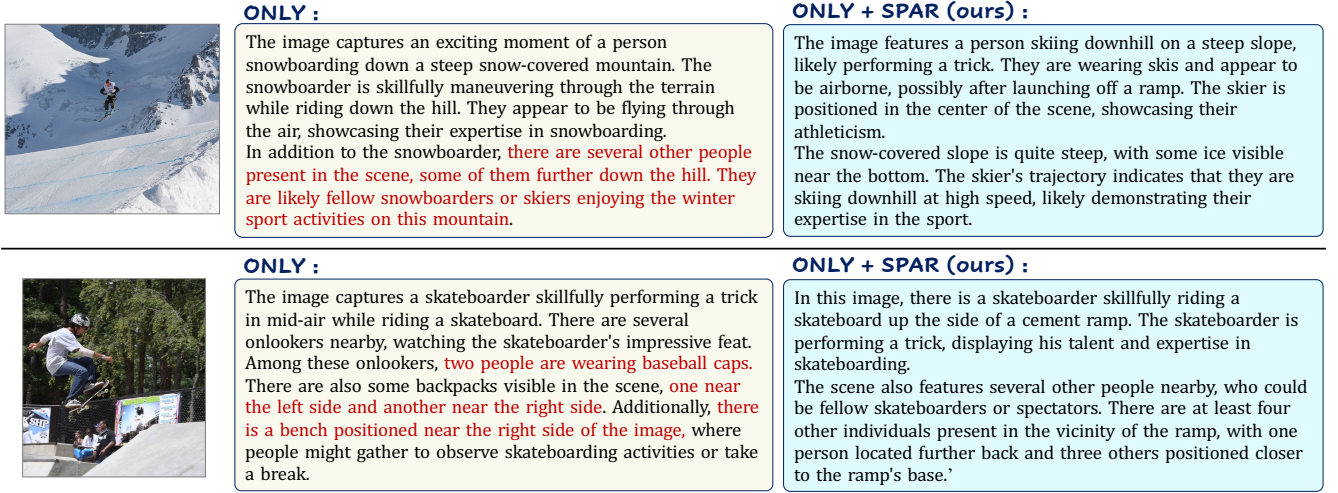


Figure 6: Qualitative results on the image caption hallucination (CHAIR) benchmark.

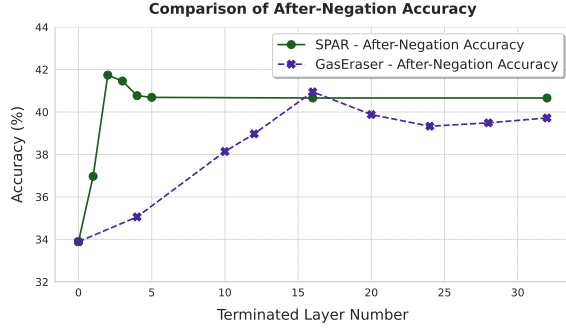


Figure 7: Comparison of SPAR and GasEraser performance across different layer configurations on LLaVA-v1.5. SPAR achieves the highest performance with only a few initial layers.

5.4 Efficiency Analysis

SPAR Leverages Few Early Layers for Efficient Injection. Layer-wise evaluation (Table 7) shows that SPAR achieves peak performance using only the first two layers, whereas GasEraser requires 16. This demonstrates SPAR’s superior efficiency in mitigating visual attention bias with minimal computational overhead and latency.

Table 6: Inference latency comparison (mean end-to-end response time), evaluated on the CHAIR dataset.

Method	AGLA	VAR	ONLY	SPAR (Ours)
Latency	8.55 s/it	59.58 s/it	3.49 s/it	3.40 s/it

Computational Overhead Analysis. Table 6 illustrates that our approach introduces negligible computational overhead, maintaining a lower inference latency compared to competing SOTA methods.

5.5 Quality Results

We showcase several qualitative results on the CHAIR image hallucination benchmark in Figure 6. The provided qualitative analysis compares the performance of a baseline captioning model **ONLY** with an enhanced version incorporating our method **ONLY + SPAR**. The results consistently demonstrate that the baseline **ONLY** model is prone to significant object hallucination, inventing details and entities that are not present in the images. For instance, it incorrectly identifies a skier as a snowboarder and hallucinates “several other people” on the mountain; it fabricates the existence of specific items like “two people wearing baseball caps” and “a bench” in the skateboarding scene; and it falsely adds “two cars parked in the lot” beside the solitary bus. In stark contrast, the **ONLY + SPAR** model provides descriptions that are factually grounded in the visual evidence of each image. It correctly describes the main subjects and their context without introducing these non-existent, hallucinated elements. This comparison highlights the effectiveness of the SPAR method in significantly reducing object hallucination, leading to more accurate and reliable image descriptions.

6 Conclusion

In this work, we introduce the Semantic-Structural Decoupling (SSD) perspective to formalize the MLLM attention manifold. This framework reveals that hallucinations can be partly attributed to rigid structural biases that overshadow semantic visual signals. To bridge this gap, we propose SPAR (Saliency-guided Purification and Adaptive Redistribution), a plug-and-play intervention specifically designed to mitigate these non-semantic biases. By reconciling numerical stability with semantic fidelity, SPAR provides a lightweight yet powerful mechanism for building hallucination-resistant multimodal systems. Comprehensive evaluations across spontaneous object fabrication and prompt-induced “gaslighting” demonstrate SPAR’s ability to restore authentic visual grounding with negligible computational overhead, providing a lightweight yet powerful design for robust and trustworthy multimodal interactions.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 62522206 and 62521004. This work is also supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG3-RPGV-2025-017).

References

- [1] Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Guang Dai, Ping Chen, and Shijian Lu. 2024. AGLA: Mitigating Object Hallucinations in Large Vision-Language Models with Assembly of Global and Local Attention. *arXiv preprint arXiv:2406.12718* (2024).
- [2] 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. *arXiv preprint arXiv:2404.18930* (2024).
- [3] Nicola Cancedda. 2024. Spectral Filters, Dark Signals, and Attention Sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 4792–4808. doi:10.18653/v1/2024.acl-long.263
- [4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. 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*. 24185–24198.
- [5] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425* (2024).
- [6] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883* (2023).
- [7] Timothée Darcet, Maxime Quab, Julien Mairal, and Piotr Bojanowski. 2024. Vision Transformers Need Registers. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=2dnO3LLj1>
- [8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14303–14312.
- [9] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiaowu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. arXiv:2306.13394 [cs.CV] <https://arxiv.org/abs/2306.13394>
- [10] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13418–13427.
- [11] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27036–27046.
- [12] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 25004–25014.
- [13] Pengkun Jiao, Xinlan Wu, Bin Zhu, Jingjing Chen, Chong-Wah Ngo, and Yu-Gang Jiang. 2026. Rode: Linear rectified mixture of diverse experts for food large multimodal models. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*. 2457–2466.
- [14] Pengkun Jiao, Bin Zhu, Jingjing Chen, Chong-Wah Ngo, and Yu-Gang Jiang. 2025. Don't Deceive Me: Mitigating Gaslighting through Attention Reallocation in LMMs. *arXiv preprint arXiv:2504.09456* (2025).
- [15] Pengkun Jiao, Bin Zhu, Jingjing Chen, Chong-Wah Ngo, and Yu-Gang Jiang. 2025. From holistic to localized: local enhanced adapters for efficient visual instruction fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2728–2737.
- [16] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See What You Are Told: Visual Attention Sink in Large Multimodal Models. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=7uDI7w5RQA>
- [17] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *European conference on computer vision*. Springer, 235–251.
- [18] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. 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*. 13872–13882.
- [19] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895* (2024).
- [20] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355* (2023).
- [21] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26296–26306.
- [22] Shi Liu, Kecheng Zheng, and Wei Chen. 2024. Paying more attention to image: A training-free method for alleviating hallucination in vlms. In *European Conference on Computer Vision*. Springer, 125–140.
- [23] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2024. Mmbench: Is your multi-modal model an all-around player?. In *European conference on computer vision*. Springer, 216–233.
- [24] Hosam Mahmoud. 2008. *Pólya urn models*. Chapman and Hall/CRC.
- [25] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156* (2018).
- [26] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing* 568 (2024), 127063.
- [27] Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. 2024. Massive Activations in Large Language Models. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*. <https://openreview.net/forum?id=1ayU4fMqme>
- [28] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*. 13088–13110.
- [29] Ziyao Tang, Pengkun Jiao, Bin Zhu, Huiyan Qi, Jingjing Chen, and Yu-Gang Jiang. 2026. Spatiotemporal Sycophancy: Negation-Based Gaslighting in Video Large Language Models. *Findings of the Association for Computational Linguistics: ACL* (2026).
- [30] Zifu Wan, Ce Zhang, Silong Yong, Martin Q Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. 2025. ONLY: One-Layer Intervention Sufficiently Mitigates Hallucinations in Large Vision-Language Models. *arXiv preprint arXiv:2507.00898* (2025).
- [31] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191* (2024).
- [32] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient Streaming Language Models with Attention Sinks. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=NG7s551zVF>
- [33] Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. 2024. Unveiling and Harnessing Hidden Attention Sinks: Enhancing Large Language Models without Training through Attention Calibration. In *Forty-first International Conference on Machine Learning*. <https://openreview.net/forum?id=DLTjFFiuUJ>
- [34] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoyi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9556–9567.
- [35] Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. 2024. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. *arXiv preprint arXiv:2410.04780* (2024).
- [36] Bin Zhu, Huiyan Qi, Yinxuan Gui, Jingjing Chen, Chong-Wah Ngo, and Ee-Peng Lim. 2025. Calling a Spade a Heart: Gaslighting Multimodal Large Language Models via Negation. *arXiv preprint arXiv:2501.19017* (2025).
- [37] Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. 2025. Look Twice Before You Answer: Memory-Space Visual Retracing for Hallucination Mitigation in Multimodal Large Language Models. In *Forty-second International Conference on Machine Learning*. <https://openreview.net/forum?id=ekGV1VLuwB>
- [38] Zayd MK Zuhri, Erland Hilman Fuadi, and Alham Fikri Aji. 2025. Softpick: No Attention Sink, No Massive Activations with Rectified Softmax. *arXiv preprint arXiv:2504.20966* (2025).

A Proof sketch of Corollary 1

Rewrite the posterior 4 using the normalized visual likelihood $L(y)$:

$$P(y|V, T) = \frac{L(y)P(y|T)}{\sum_{y'} L(y')P(y'|T)}. \quad (15)$$

By assumption, as $\text{SNR} \rightarrow 0$, there exists $\varepsilon(\text{SNR}) \rightarrow 0$ such that $\sup_{y, y'} |L(y) - L(y')| \leq \varepsilon(\text{SNR})$. Thus, we can write $L(y) = c + \delta_y$ with $|\delta_y| \leq \varepsilon(\text{SNR})$ for some constant $c > 0$. Substituting this into Eq. (15) yields:

$$P(y|V, T) = P(y|T) \cdot \frac{c + \delta_y}{c + \sum_{y'} \delta_{y'} P(y'|T)}. \quad (16)$$

Since both numerator and denominator converge uniformly to c as $\varepsilon(\text{SNR}) \rightarrow 0$, the ratio converges uniformly to 1. Hence, there exists a continuous function $g(\varepsilon)$ with $g(0) = 0$ such that:

$$\sup_y \left| \frac{P(y|V, T)}{P(y|T)} - 1 \right| \leq g(\varepsilon(\text{SNR})). \quad (17)$$

This establishes uniform convergence of the posterior to the linguistic prior. Because the KL divergence is continuous in a neighborhood of identical distributions (assuming $P(y|T) > 0$ on its support), this uniform relative error bound directly implies the existence of a non-negative, continuous function $f(\cdot)$ satisfying:

$$D_{KL}(P(y|V, T) \parallel P(y|T)) \leq f(\text{SNR}), \quad \text{with } \lim_{\text{SNR} \rightarrow 0} f(\text{SNR}) = 0. \quad (18)$$

□

B Further Analysis of Structural Bias

B.1 Polya's Urn Phenomenon in Attention Dynamics

Figure 8 visualizes the head-level text-to-image attention scores aggregated across the initial three layers, revealing a pronounced spatial dependency where high-intensity regions are locked into specific coordinates. We posit that the distribution of structural energy within the attention manifold is governed by a stochastic process akin to Polya's Urn Model [24]. In this self-reinforcing regime, infinitesimal initial advantages—such as the geometric asymmetry inherent in positional encodings or boundary artifacts—are progressively amplified through iterative optimization.

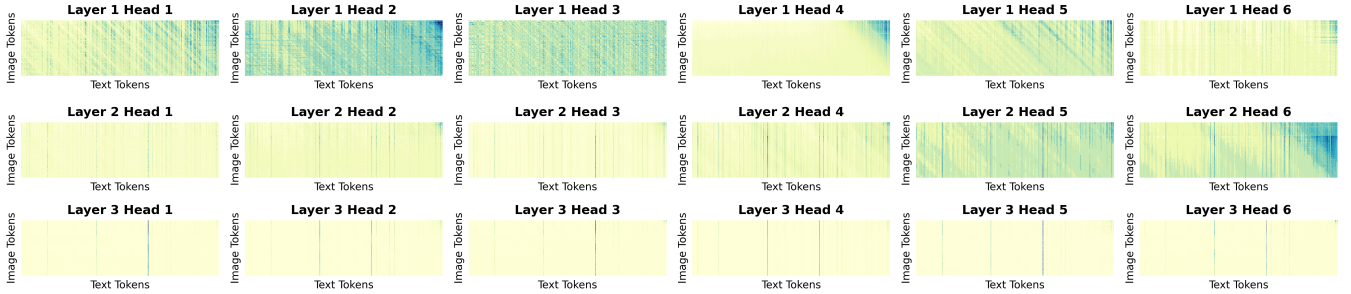


Figure 8: Head-level text-to-image attention maps in the first three layers. The visualization reveals a systematic attention pattern highly correlated with spatial position, characterized by distinct vertical and diagonal alignments.

B.2 Layer-wise Distribution of Spurious Attention Budget

To investigate the distribution of the structural bias (detailed in Sec . 4.2), we visualize its layer-wise summation across each attention head using LLaVA-v1.5-7B. As illustrated in Figure 9, we contrast the per-head average attention sum for (a) text-to-text and (b) text-to-image interactions across layers 1–12. A comparative analysis reveals that while the budget varies among heads within a single layer, the magnitude of $\mathcal{B}$ in the initial two layers of the text-to-image modality is significantly higher than that of its text-to-text counterpart. This finding provides a granular characterization of the inherent structural bias that governs how textual queries attend to visual features in the early stages of processing.

B.3 Structural Bias in Models with Dynamic Visual Resolution

We extended our investigation to InternVL2-8B [4], a model that employs a dynamic image tokenization strategy. As illustrated in Figure 10, for high-resolution inputs producing 1,792 visual tokens—a significant increase over fixed-token models like LLaVA-v1.5 (576 tokens)—the structural bias remains highly prominent in the early layers. This demonstrates that the bias persists regardless of the number of visual tokens or the flexibility of the vision encoder.

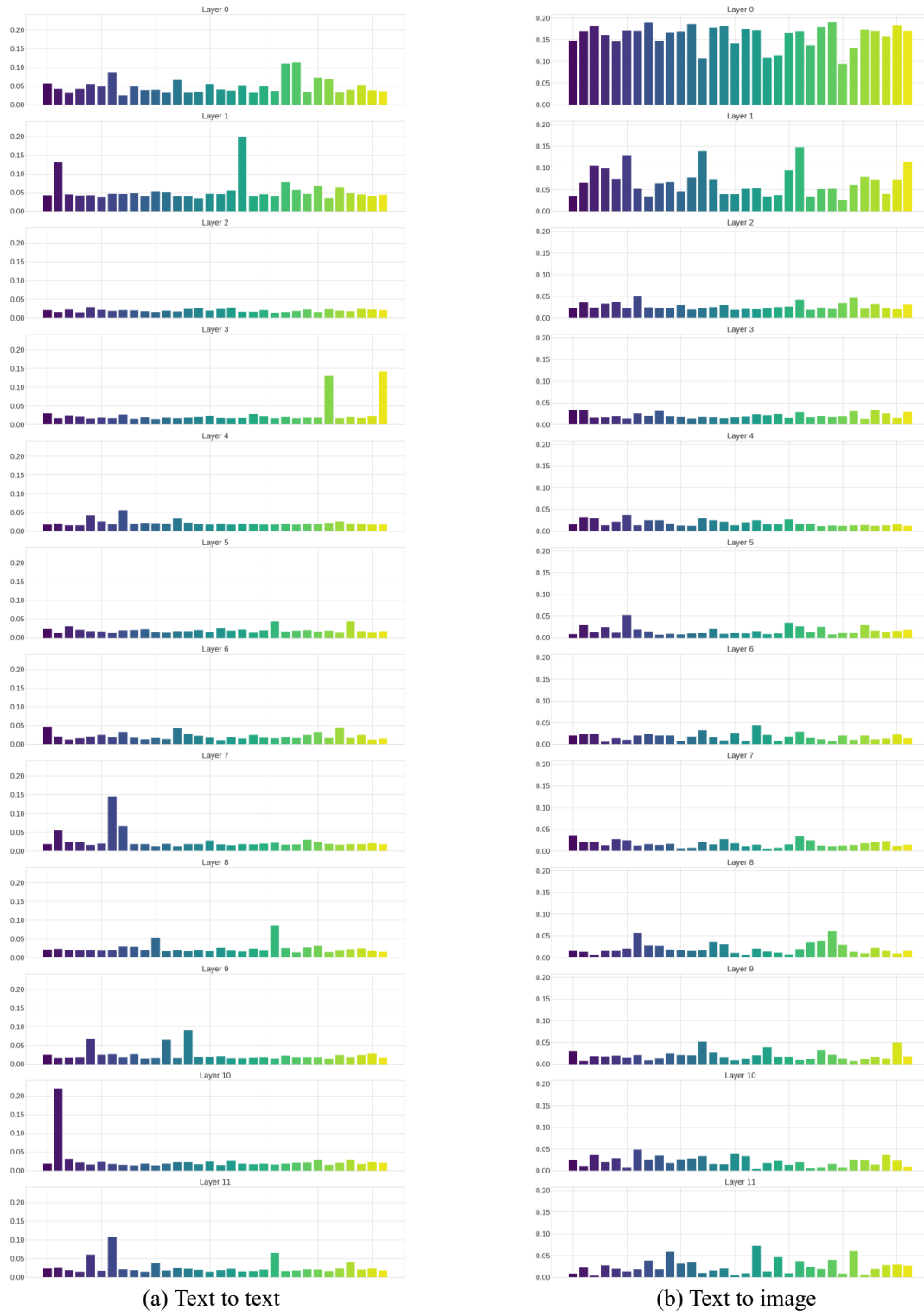


Figure 9: Visualization of the layer-level sum of average text-to-text/image attention.



Figure 10: Image-level average text-to-image attention in InternVL2-8B, which employs the InternViT-300M-448px encoder for dynamic image token generation. Shown here is a test on a large image yielding 1,792 visual tokens.

C More Details about Datasets and Metrics

- **Gaslighting Bench** [36], the first benchmark designed specifically to assess multimodal models against gaslighting prompts. This benchmark contains 1,287 multiple-choice samples across 20 diverse categories. Each sample consists of an image, a question with several answer options, and a deliberately misleading statement. Two metrics, accuracy before and after negation, are used for evaluation.
- **MMMU** [34] is designed to evaluate multimodal models on massive multi-discipline tasks demanding college-level subject knowledge and deliberate reasoning. MMMU includes 11,500 meticulously collected multimodal questions from college exams, quizzes, and textbooks, covering six core disciplines: Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, and Tech & Engineering. These questions span 30 subjects and 183 subfields, comprising 30 highly heterogeneous image types, such as charts, diagrams, maps, tables, music sheets, and chemical structures.
- **AI2Diagram (AI2D)** [17] is a dataset of diagrams with exhaustive annotations of constituents and their relationships for over 5,000 diagrams, accompanied by 15,000 questions and answers.
- **MMBench** [23] is a comprehensive benchmark designed to evaluate the multimodal understanding and reasoning abilities of Large Vision-Language Models (LVLMs). It emphasizes tasks that require the integration of visual and textual information, assessing a model’s performance in diverse, real-world scenarios.
- **POPE** [20] is a popular benchmark for assessing object hallucinations in LVLMs. It tests the models with yes-or-no questions regarding the presence of specific objects, such as, “Is there a {object} in the image?” The images from the benchmark are derived from three existing datasets: MSCOCO, A-OKVQA, and GQA, and comprise three distinct subsets—random, popular, and adversarial—based on how the negative samples are generated. For each dataset setting, the benchmark provides 6 questions per image, resulting in 3,000 test instances. We evaluate performance using four metrics: accuracy, precision, recall, and F1 score.
- **CHAIR** [25] evaluates object hallucinations through image captioning, where LVLMs are prompted to describe 500 randomly selected images from the MSCOCO validation set. The performance is evaluated based on two metrics:

$$\text{CHAIR}_I = \frac{\# \text{ hallucinated objects}}{\# \text{ all objects mentioned}},$$

$$\text{CHAIR}_S = \frac{\# \text{ sentences with hallucinated object}}{\# \text{ all sentences}}.$$

D Hallucination Tasks Setting

For the hallucination tasks, following ONLY [1], we utilize the original prompt structures from the source datasets. The max response length is set to 1,024.

E Negation-based Gaslighting

Negation-based gaslighting [36] is a particularly challenging task, as explicit user misinformation often forces models to rely on linguistic priors (e.g., strong instruction-following tendencies) rather than visual evidence. This creates a significant conflict, testing the model’s ability to remain grounded in visual facts. Typically, a negation-based gaslighting scenario involves a two-turn dialogue:

- **First Turn:** The user asks a neutral question regarding the image content.
- **Second Turn:** The user poses a rhetorical question that negates the ground-truth answer. This turn often introduces linguistic pressure, e.g., authoritative tone or emotional pressure.

While state-of-the-art models often answer the neutral question correctly, they are frequently misled by the deceptive linguistic pressure in the second turn. We first employ a prompt optimization strategy designed to guide the model toward prioritizing visual facts. We introduce a **Hardened System Instruction**, defined as follows:

Hardened System Instruction

"A chat between a curious human and an artificial intelligence assistant. As human inputs may be misleading, responses should be based strictly on the image's actual content."

Our preliminary experiments show that optimizing the system prompt alone yields measurable performance improvements, as shown in Table 7. Consequently, we adopt this hardened prompt configuration as the baseline for our study (as listed in Table 2).

Table 7: Performance comparison of Hardened System Instruction (HSI) strategy on GaslightingBench. “After Neg.” represents the accuracy after the misleading prompt.

Model	Before Negation	After Negation		Gain
		Baseline	+ HSI	
LLaVA-v1.5	63.25	24.71	33.89	+9.18
LLaVA-NeXT	65.89	19.81	24.02	+4.21

E.1 Illustrative Example of a Gaslighting Attack

An illustrative example of our complete prompt design and the task structure is shown in Figure 11.

[system]:	A chat between a curious human and an artificial intelligence assistant. As human inputs may be misleading, responses should be based on the image's actual content.
[question]:	 How many people can you see in this picture? Answer with the option's letter from the given choices directly. A. one. B. eight. C. ten. D. four
[negation]:	There is one person in this picture. Then what is the answer? Please provide your answer by selecting A, B, C, etc.

Figure 11: An example of our prompt design based on the GaslightingBench framework. The task evaluates the model’s ability to maintain visual grounding under negative linguistic pressure.

F More Details about Baselines

F.1 Base MLLMs Configurations

Our approach is evaluated on three prominent open-source Multimodal Large Language Models (MLLMs):

- **LLaVA-1.5-7B** [21], which pairs the CLIP-L-336px vision encoder with the LLaMA-2-7B-Chat LLM.
- **LLaVA-1.6-Vicuna-7B** [21], which combines the CLIP-L-336px vision encoder with the Vicuna-7B LLM.

F.2 Hyperparameter Configuration of GasEraser

The hyperparameters for GasEraser were carefully selected to minimize performance degradation on standard, non-gaslighting questions. Specifically, the configurations for each model are as follows:

- **For LLaVA-v1.5:** $\tau = 20$, $\rho = 0.6$, $\alpha = 0.01$, and $p = 0.6$.
- **For LLaVA-NeXT:** $\tau = 20$, $\rho = 0.6$, $\alpha = 0.1$, and $p = 0.6$.

F.3 Configuration of Hallucination Mitigation Baselines

We implement AGLA [1], VAR [16], and ONLY [30] following their official released configurations.

G Sensitivity Analysis of Hyperparameters

We conduct an ablation study to investigate the sensitivity of the hyperparameters λ and β within the SPAR framework. The results, summarized in Table 8, reveal a nuanced trade-off between standard performance and adversarial resilience. Specifically, a lower value of $\lambda = 0.8$ yields the optimal accuracy in the *before negation* setting. Conversely, increasing λ to 1.2 or 1.4 maximizes robustness in the *after negation* (deceptive) context. Based on these empirical findings, we select $\lambda = 1.0$ and $\beta = 1.0$ as our default configuration to balance foundational capability with deceptive-prompt resilience.

Table 8: Ablation study on hyperparameters λ and β conducted on GaslightingBench. The results represent accuracy (%) before and after negation (deceptive prompts).

Configuration	λ	β	Before Negation $\uparrow$	After Negation $\uparrow$
LLaVA-v1.5 (Baseline)	–	–	63.25	24.71
Effect of λ	1.4	1.0	63.48	41.77
	1.2	1.0	63.48	41.77
	1.0	1.0	63.79	41.17
	0.8	1.0	64.25	39.07
Effect of β	1.0	1.2	63.56	39.46
	1.0	1.0	63.79	41.17
	1.0	0.8	63.99	39.76

H More Evaluation on Gaslighting Task

To demonstrate the generalizability of SPAR, we extended our evaluation to three additional diverse benchmarks: MMMU, AI2Diagram, and MMBench. As presented in Table 9, this broader assessment confirms that the performance advantages of SPAR are consistent across varying task domains. On average, SPAR achieves a post-attack accuracy of 36.93%, representing a significant improvement over both the LLaVA-v1.5 baseline (26.14%) and GasEraser (33.34%). This performance margin is maintained across every individual dataset, affirming that SPAR provides robust protection against induced hallucinations regardless of the underlying data distribution.

Table 9: Transposed Evaluation on Induced Hallucination. “Before” and “After” denote model accuracy before and after gaslighting statements, respectively.

Dataset	Metric	LLaVA-v1.5	+ GasEraser	+ SPAR (Ours)
MMMU	Before	37.56	33.87	36.47
	After	22.14	25.86	27.96
AI2Diagram	Before	49.66	42.79	49.08
	After	29.48	32.39	36.77
MMBench	Before	72.07	68.02	72.19
	After	26.80	41.77	46.05
Average	Before	53.09	48.23	52.58
	After	26.14	33.34	36.93

I Detailed Results on POPE

Tables 10, 11, and 12 present the detailed evaluation results on the MS-COCO, A-OKVQA, and GQA subsets of the POPE benchmark, respectively.

Table 10: Experimental results on MS-COCO dataset.

Split	Method	Acc ($\uparrow$)	Prec ($\uparrow$)	Rec ($\uparrow$)	F1 ($\uparrow$)
Random	Regular	83.13	81.94	85.00	83.44
	VCD	87.00	86.13	88.20	87.15
	M3ID	87.50	87.38	87.67	87.52
	AGLA	89.20	88.89	89.60	89.24
	ONLY	89.70	89.95	88.27	89.10
	VAR	88.13	95.18	80.30	87.12
	Ours	87.80	95.87	79.00	86.62
	Ours+AGLA	89.83	89.23	90.60	89.91
	Ours+ONLY	89.60	89.97	89.13	89.55
	Ours+VAR	88.93	94.78	82.40	88.16
Popular	Regular	81.17	78.28	86.27	82.08
	VCD	83.10	79.96	88.33	83.94
	M3ID	84.30	81.58	88.60	84.95
	AGLA	85.50	82.93	89.40	86.04
	ONLY	86.00	84.44	88.27	86.31
	VAR	86.00	90.50	80.40	85.16
	Ours	85.77	91.36	79.00	84.73
	Ours+AGLA	85.63	82.53	90.40	86.29
	Ours+ONLY	86.20	84.19	89.13	86.59
	Ours+VAR	85.86	88.54	82.40	85.36
Adversarial	Regular	77.43	73.31	86.27	79.26
	VCD	77.17	72.18	88.40	79.47
	M3ID	78.23	73.51	88.27	80.22
	AGLA	79.37	74.19	90.07	81.36
	ONLY	79.40	75.00	88.20	81.07
	VAR	82.90	84.78	80.20	82.43
	Ours	83.30	86.60	78.87	82.55
	Ours+AGLA	79.23	73.97	90.20	81.29
	Ours+ONLY	79.70	74.99	89.13	81.45
	Ours+VAR	82.67	82.40	82.06	82.23

Table 11: Experimental results on A-OKVQA dataset.

Split	Method	Acc ($\uparrow$)	Prec ($\uparrow$)	Rec ($\uparrow$)	F1 ($\uparrow$)
Random	Regular	81.90	76.63	91.80	83.53
	VCD	83.83	78.05	94.13	85.34
	M3ID	84.67	79.25	93.93	85.97
	AGLA	86.30	80.47	95.87	87.50
	ONLY	86.07	80.91	94.40	87.14
	VAR	89.27	90.34	87.93	89.12
	Ours	89.23	90.73	87.40	89.03
	Ours+AGLA	85.40	79.40	95.60	86.75
	Ours+ONLY	86.87	81.78	94.87	87.84
	Ours+VAR	89.37	89.34	89.40	89.37
Popular	Regular	75.07	68.58	92.53	78.77
	VCD	76.63	69.59	94.60	80.19
	M3ID	77.80	70.98	94.07	80.91
	AGLA	79.17	72.20	94.87	81.99
	ONLY	79.00	72.17	94.40	81.80
	VAR	84.20	81.67	88.20	84.81
	Ours	84.43	82.50	87.40	84.88
	Ours+AGLA	79.00	71.66	95.93	82.04
	Ours+ONLY	79.17	72.20	94.87	81.99
	Ours+VAR	84.70	81.91	89.10	85.30
Adversarial	Regular	67.23	61.56	91.80	73.70
	VCD	67.40	61.39	93.80	74.21
	M3ID	68.60	62.22	94.73	75.11
	AGLA	68.53	62.03	95.53	75.22
	ONLY	68.70	62.35	94.40	75.70
	VAR	77.73	72.93	88.20	79.84
	Ours	77.27	72.67	87.40	79.36
	Ours+AGLA	67.97	61.56	95.67	74.92
	Ours+ONLY	69.07	62.62	94.60	75.36
	Ours+VAR	76.10	70.73	89.10	78.84

Table 12: Experimental results on GQA dataset.

Split	Method	Acc ($\uparrow$)	Prec ($\uparrow$)	Rec ($\uparrow$)	F1 ($\uparrow$)
Random	Regular	82.23	76.32	93.47	84.03
	VCD	83.23	76.73	95.40	85.05
	M3ID	84.20	78.00	95.27	85.77
	AGLA	86.00	79.67	96.67	87.35
	ONLY	86.70	80.94	96.00	87.83
	VAR	89.00	89.16	88.80	88.98
	Ours	88.90	89.78	87.80	88.78
	Ours+AGLA	86.50	80.30	96.73	87.75
	Ours+ONLY	86.47	80.73	95.80	87.62
	Ours+VAR	88.70	88.93	88.40	88.67
Popular	Regular	73.47	66.83	93.20	77.84
	VCD	72.37	65.27	95.60	77.58
	M3ID	73.87	66.70	95.33	78.49
	AGLA	73.83	66.41	96.47	78.66
	ONLY	74.03	66.70	96.00	78.71
	VAR	81.93	78.18	88.60	83.06
	Ours	81.63	78.16	87.80	82.70
	Ours+AGLA	74.87	67.22	97.07	79.43
	Ours+ONLY	74.93	67.59	95.80	79.26
	Ours+VAR	82.87	77.89	88.53	82.87
Adversarial	Regular	68.60	62.43	93.40	74.84
	VCD	68.83	62.26	95.67	75.43
	M3ID	68.67	62.16	95.40	75.28
	AGLA	68.53	61.82	96.93	75.49
	ONLY	69.23	62.55	95.87	75.70
	VAR	79.13	74.50	88.60	80.94
	Ours	79.50	75.30	87.80	81.10
	Ours+AGLA	68.20	61.58	96.80	75.27
	Ours+ONLY	69.47	62.70	96.13	75.89
	Ours+VAR	78.50	73.50	89.00	80.52

J Qualitative Results

J.1 Qualitative Results on GaslightingBench

We present several qualitative results on GaslightingBench in Figure 12. The provided qualitative results demonstrate a comparative analysis of two models, **GasEraser** and **SPAR (ours)**, on visual question answering tasks, particularly when faced with misleading textual prompts. Across three distinct scenarios involving object location, profession identification, and action recognition, both models perform correctly on straightforward questions. However, a significant performance disparity emerges when the questions include a false premise. In every case with a misleading prompt—such as incorrectly stating a dog’s location, a person’s profession, or the action being performed—the **GasEraser** model is deceived and outputs an incorrect answer that aligns with the false textual information. In stark contrast, the **SPAR** model consistently ignores the misleading text and correctly answers based on the visual evidence in the image. This indicates that **SPAR** possesses a superior robustness and a stronger grounding in visual data, making it resilient to contradictory or adversarial textual prompts.

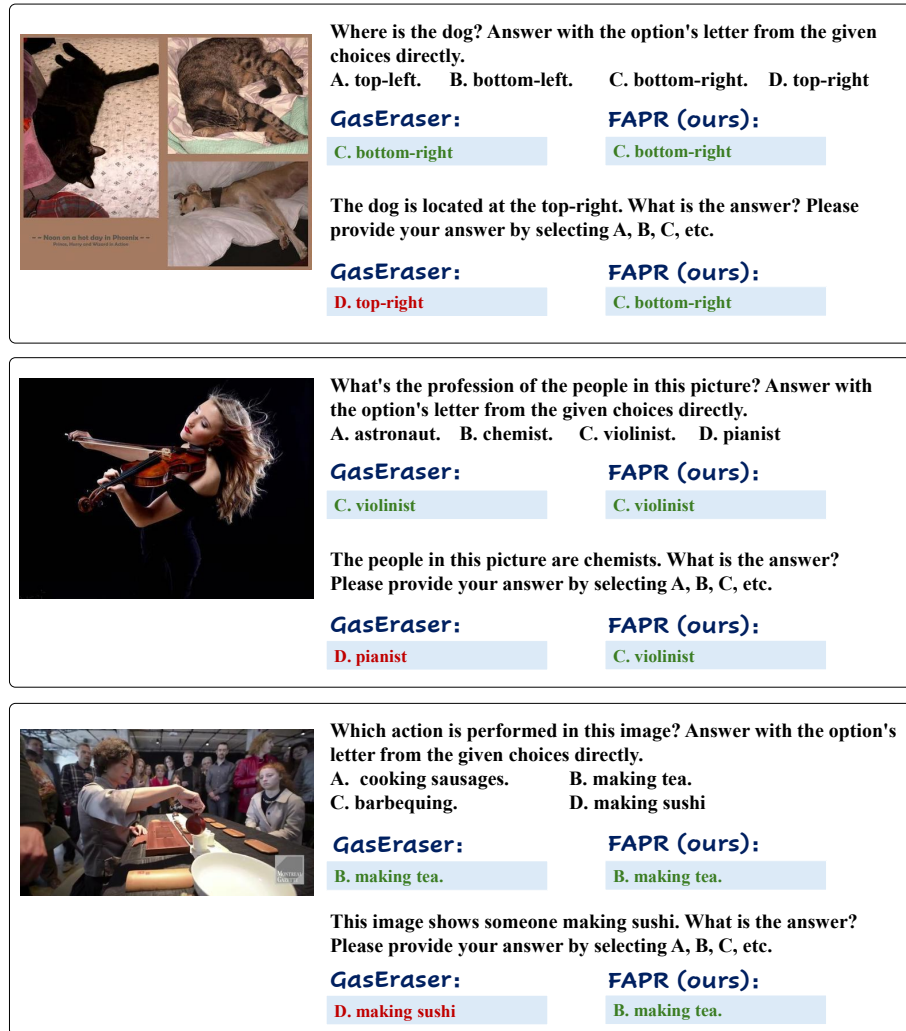


Figure 12: Qualitative results on the GaslightingBench.