Vision-Language Introspection: Mitigating Overconfident Hallucinations in MLLMs via Interpretable Bi-Causal Steering

Anonymous ACL submission

Abstract

Object hallucination critically undermines the reliability of Multimodal Large Language Models, often stemming from a fundamental failure in cognitive introspection—where models blindly trust linguistic priors over specific visual evidence. Existing mitigations remain limited: contrastive decoding approaches operate superficially without rectifying internal semantic misalignments, while current latent steering methods rely on static vectors that lack instance-specific precision. We introduce **Vision-Language Introspection (VLI)**, a training-free inference framework that simulates a metacognitive self-correction process. VLI first performs *Attributive Introspection* to diagnose hallucination risks via probabilistic conflict detection and localize the causal visual anchors. It then employs *Interpretable Bi-Causal Steering* to actively modulate the inference process, dynamically isolating visual evidence from background noise while neutralizing blind confidence through adaptive calibration. VLI achieves state-of-the-art performance on advanced models, reducing object hallucination rates by 12.67% on MMHal-Bench and improving accuracy by 5.8% on POPE.

1 Introduction

Multimodal Large Language Models (MLLMs) have advanced significantly in reasoning but suffer critically from object hallucination, generating plausible yet non-existent objects (Liu et al., 2024a). Recent studies identify this not merely as perceptual error, but a failure of *cognitive introspection*: models exhibit blind confidence, over-relying on linguistic priors rather than verifying generation against specific visual evidence (Min et al., 2024; Zhou et al., 2023).

Current mitigation strategies have shifted from costly retraining (Ding et al., 2025; Jiang et al., 2024; Xing et al., 2024) toward lightweight training-free paradigms (Chen et al., 2025d; Favero et al., 2024; Wu et al., 2025a; Yin et al., 2024; Zhang et al., 2025), broadly bifurcating into distribution-level and representation-level interventions. However, both paradigms face fundamental limitations in *precision* and *cognitive depth*. Distribution-level methods, like Contrastive Decoding (Leng et al., 2024), often indiscriminately mask visual inputs (An et al., 2025; Chen et al., 2024), inadvertently discarding background context essential for reasoning (Fu et al., 2025; Zhao et al., 2025a) and failing to rectify deep-seated erroneous visual-semantic connections (Liu et al., 2024b; He et al., 2025). Conversely, representation-level strategies relying on static steering vectors (Shi et al., 2025; Li et al., 2025b; Suo et al., 2025) lack instance-specific granularity (Liu et al., 2025; Su et al., 2025; Yang et al., 2025). Crucially, they fail to address the model's intrinsic *blind confidence* (Duan et al., 2025; Ye et al., 2025), as generic interventions cannot decouple true visual understanding from stubborn hallucinatory biases (Zhu et al., 2025b; Kalai et al., 2025).

Crucially, existing methods treat multimodal reasoning as a binary selection between linguistic priors and visual features, overlooking the cognitive necessity of **explicit causal dependency**. Analogous to human perception, where expectations are actively verified against specific visual regions, true reasoning demands dynamic self-verification rather than static probability ranking. Current LVLMs lack this capability, causing blind confidence. While mechanistic interpretability diagnoses these states (Jiang et al., 2025), we argue that diagnosis must be operationalized into active control to bridge the gap between passive interpretation and rectification (Park et al., 2025; Chen et al., 2025a; Bae et al., 2025).

Building on these insights, we propose **Vision-Language Introspection (VLI)**, a training-free framework simulating metacognitive self-correction. Unlike post-hoc methods (Chen et al.,

2025b; Heiman et al., 2025), VLI employs a bidirectional mechanism aligning visual evidence with textual generation. First, *Attributive Introspection* uses attention purification to isolate the causal visual anchor, strictly differentiating object pixels from background context (Zhao et al., 2024). Second, *Interpretable Bi-Causal Steering* rectifies inference by constructing **Anchor-Only** and **Context-Only** counterfactual states via inpainting. This derives a dynamic correction vector that enhances focus on visual evidence while suppressing background noise triggering linguistic priors. Finally, *Adaptive Confidence Calibration* (Xie et al., 2024) addresses blind confidence by measuring cognitive conflict between holistic and counterfactual states, adaptively penalizing ungrounded certainty.

Extensive experiments demonstrate that VLI significantly outperforms the baseline by reducing MMHal hallucination rates by up to 12.67% and enhancing POPE accuracy by 6.33%, while surpassing state-of-the-art methods by margins of 5.37% and 1.60%, respectively. Our contributions are:

- We propose **VLI**, a framework that systematically diagnoses and rectifies object hallucinations by interpreting and manipulating specific visual anchors.
- We introduce **Bi-Causal Steering**, which performs precise latent interventions by dynamically contrasting Anchor-Only vs. Context-Only representations to isolate and reinforce the true visual cause.
- We integrate **Adaptive Confidence Calibration** to detect cognitive conflict during inference, preventing hallucinations driven by blind confidence.

2 Related Work

Hallucination mitigation strategies for Multimodal Large Language Models generally categorize into training-based alignment and training-free inference intervention. While training-based methods align cognition via RLHF (Sun et al., 2024) or DPO (Yang et al.), we focus on the training-free paradigm to avoid prohibitive costs (Chen et al., 2025b; Ding et al., 2025; Fu et al., 2025). These approaches typically involve surface-level decoding manipulation (Jiang et al., 2024; Min et al., 2024; Ren et al., 2025) or deep latent state intervention (Wu et al., 2025b,a; Xiao et al., 2025).

2.1 Inference-Time Hallucination Mitigation

Decoding Strategies. This stream rectifies hallucinations by externally calibrating output probabilities. *Contrastive Decoding* mitigates linguistic priors by contrasting original logits against distorted ones (VCD (Leng et al., 2024)) or utilizing decoupled projectors (IBD (Zhu et al., 2025a), DCD (Chen et al., 2025d)). Recent works like CICD (Zhao et al., 2025c) and DeGF (Zhang et al., 2025) employ cross-image references or generative feedback (An et al., 2025; Lee and Song, 2025) to preserve visual details (Lyu et al., 2024; Shi et al., 2025). To improve flexibility, *Adaptive Strategies* such as Octopus (Suo et al., 2025) and MoD (Chen et al., 2025e) dynamically route decoding strategies, while DLC (Chen et al., 2025c) and M3ID (Favero et al., 2024) perform real-time logit calibration. Additionally, *Penalty-Based Mechanisms* like OPERA (Huang et al., 2024) and DOPRA (Wei and Zhang, 2024) modify beam search to penalize overtrust patterns (Li et al., 2025a). However, these decoding methods function primarily as surface-level regularizers, failing to rectify the corrupted internal representations (Kaul et al., 2024).

Latent Space Steering. A more intrinsic paradigm directly modulates hidden states. Methods such as VTI (Liu et al., 2025) and Nullu (Yang et al., 2025) employ global steering vectors derived from feature averaging or null-space projection. Others like VaLSe (Chen et al., 2025a) and Truth-PrInt (Duan et al., 2025) use probes for guided intervention (Park et al., 2025). Unlike these approaches that rely on static, dataset-level vectors, our VLI framework introduces *Interpretable Bi-Causal Steering*. We compute a precise, dynamic steering vector derived from the cognitive gap between counterfactual states, enabling the model to actively introspect and correct instance-specific visual grounding.

2.2 Mechanistic Interpretability for Visual Grounding

Our work is grounded in mechanistic interpretability, which diagnoses internal attention allocation. Studies reveal that LVLMs rely on specific expert heads for semantic tracking (Zhao et al., 2025b; Deng and Yang, 2025) but often suffer from visual attention sinks (Kang et al., 2025). While existing methods utilize these findings for passive analysis or re-weighting (Liu et al., 2024b; Jiang et al., 2025; Tang et al., 2025), VLI operationalizes them

into active control. By employing *Attributive Introspection*, we extract pixel-precise causal anchors to bridge the gap between passive diagnosis and active error rectification.

3 Methodology

We introduce **Vision-Language Introspection (VLI)**, a training-free inference-time framework designed to mitigate overconfident hallucinations, illustrated in Fig. 1. Unlike standard decoding interventions that passively suppress likely tokens, VLI simulates a metacognitive self-verification process. It addresses the fundamental disconnect between linguistic priors and visual evidence through a bidirectional mechanism: 1) **Attributive Introspection**, a diagnostic phase that traces high-risk predictions back to their causal visual origins; and 2) **Bi-Causal Steering**, an intervention phase that dynamically isolates the specific visual evidence from background noise to rectify latent representations across all model layers and calibrate blind confidence. Both process is analized with Case Study in Appendix F.

3.1 Attributive Introspection: Causal Source Localization

During the generation of the t-th token, the objective of this phase is to introspect the model's reasoning process. We aim to trace the cognitive dissonance between the model's internal priors and the actual visual input back to specific image regions, formalizing this causal origin as a pixel-precise source anchor mask $\mathcal{M}_s$.

3.1.1 Introspective Conflict Detection

To quantify hallucination risks, we evaluate the consistency between visual evidence and linguistic priors via a comparative analysis of internal model states. We define two parallel decoding paths at time step t, tracking hidden states across all L layers.

First, the **Grounded Path** represents the standard reasoning process where the model receives complete visual features $V = \mathcal{E}_V(I)$ and textual context $T_{<t}$. For each layer $l \in \{1, \dots, L\}$:

$$h_{g,l}^{(t)} = \mathcal{F}_{\text{VLM}}^{(l)}(V \oplus \mathcal{E}_T(T_{<t}), h_{g,l-1}^{(t)}) \quad (1)$$

where $\mathcal{F}_{\text{VLM}}^{(l)}$ denotes the operation of the l-th transformer layer, and $h_{g,L}^{(t)}$ is the final latent state. The probability distribution $P_g^{(t)}$ over the vocabulary $\mathcal{V}$ is obtained via a linear projection W_o followed by a Softmax operation σ:

$$P_g^{(t)} = \sigma(W_o h_{g,L}^{(t)}) \quad (2)$$

Second, the **Ungrounded Path** simulates reasoning relying solely on linguistic priors by masking the visual input (replacing V with null tokens $\emptyset$), yielding hidden states $h_{u,l}^{(t)}$ and distribution $P_u^{(t)}$:

$$P_u^{(t)} = \sigma(W_o \mathcal{F}_{\text{VLM}}^{(l)}(\emptyset \oplus \mathcal{E}_T(T_{<t}), h_{u,l-1}^{(t)})) \quad (3)$$

We posit that the divergence between these two paths, termed **Introspective Conflict**, serves as a robust proxy for hallucination. We calculate the hallucination risk score $\mathcal{C}$ using the Jensen-Shannon (JS) divergence:

$$\mathcal{C}(P_g^{(t)}, P_u^{(t)}) = H\left(\frac{P_g^{(t)} + P_u^{(t)}}{2}\right) - \frac{1}{2}[H(P_g^{(t)}) + H(P_u^{(t)})] \quad (4)$$

where $H(P)$ is the Shannon entropy. Upon detecting high risk (i.e., $\mathcal{C}$ exceeds a predefined threshold θ, which is analyzed in Sec 4.5) and Appendix F.1, we identify the most suspicious token t_s. This token is defined as the vocabulary item exhibiting the maximal logarithmic divergence between the grounded and ungrounded probabilities:

$$t_s = \underset{w \in \mathcal{V}}{\text{argmax}}(\log P_g^{(t)}(w) - \log P_u^{(t)}(w)) \quad (5)$$

3.1.2 Causal Attention Purification

Having identified the locus of conflict t_s, we aim to eliminate systemic biases to extract the authentic visual evidence supporting t_s. We focus on identifying reliable attention heads to filter noise. Inspired by SEVI (Zhao et al., 2025b), we posit that only a subset of expert heads maintain reliable causal links between semantics and visual regions. We perform offline calibration on a validation set $\mathcal{D}_{val}$. Let $\mathbf{A}_{l,h}(t_s, V) \in \mathbb{R}^{N_v}$ be the attention distribution of the h-th head in the l-th layer over visual tokens V when generating t_s. We define the *localization accuracy score* $\mu_{l,h}$ as the expected probability mass falling within the ground-truth region $R_{gt}(t_s)$:

$$\mu_{l,h} = \mathbb{E}_{(I,t_s)\sim\mathcal{D}_{val}}\left[\sum_{j=1}^{N_v} \mathbf{A}_{l,h}(t_s)_j \cdot \mathbb{I}(v_j \in R_{gt}(t_s))\right] \quad (6)$$

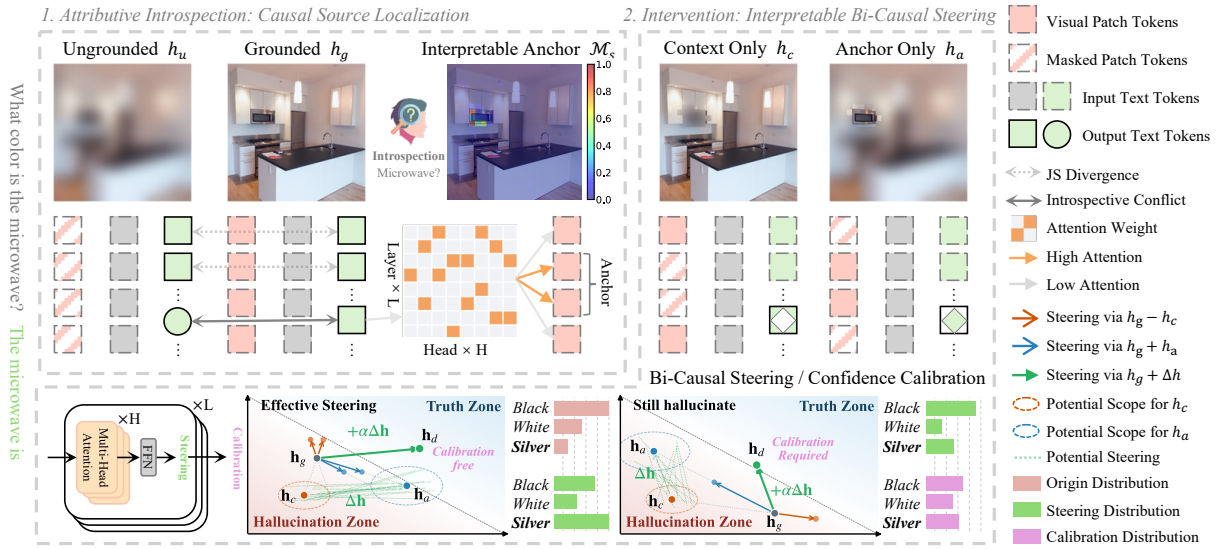

Figure 1: Overview of **VLI** framework. VLI first detects *Introspective Conflict* (§3.1.1) between grounded (h_g) and ungrounded (h_u) paths to localize the causal anchor $\mathcal{M}_s$ (§3.1.3) via purified expert attention (§3.1.2, Fig. 3). It then applies *Bi-Causal Steering* (§3.2.2) using the robust difference vector ($h_a - h_c$), which counters the scope instability of individual counterfactual states (Fig. 4). Finally, *Adaptive Confidence Calibration* (§3.2.3) penalizes blind confidence to mitigate persistent hallucinations.

We select the top M heads maximizing $\mu_{l,h}$ to construct the expert set $\mathcal{H}_{expert}$, following (Zhao et al., 2025b). To obtain the purified attention map, we aggregate the attention weights solely from these expert heads, effectively suppressing noise and attention sinks (see Appendix C for robustness analysis against visual sinks). The unnormalized purified heatmap $\tilde{H}_{att}$ is calculated as:

$$\tilde{H}_{att}(t_s) = \sum_{(l,h)\in\mathcal{H}_{expert}} \mathbf{A}_{l,h}(t_s, V) \qquad (7)$$

The final normalized attention distribution is $H_{att}(t_s) = \frac{\tilde{H}_{att}(t_s)}{\sum_{j=1}^{N_v} \tilde{H}_{att}(t_s)_j}$.

3.1.3 Interpretable Anchor Extraction

To accommodate the diversity of attention distributions, we employ a **Cumulative Energy Thresholding** strategy controlled by a single hyperparameter ρ (set to 0.4, which is analyzed in Sec 4.5). Let $\mathbf{h}_{sorted}$ denote the flattened and descendingly sorted vector of $H_{att}(t_s)$. We identify the minimal set of top-ranking pixels required to capture a total energy proportion of ρ:

$$k = \underset{i}{\operatorname{argmin}} \sum_{j=1}^{i} \mathbf{h}_{sorted}[j] \geq \rho \cdot \sum_{n=1}^{N_v} \mathbf{h}_{sorted}[n] \qquad (8)$$

The final binary causal mask is generated by selecting these top-k pixels: $\mathcal{M}_s = \mathbb{I}(H_{att}(t_s) \geq \mathbf{h}_{sorted}[k])$. This method adaptively locks onto the visual regions constituting the primary semantics based on energy concentration, effectively filtering long-tail noise without requiring a pixel count.

3.2 Intervention: Interpretable Bi-Causal Steering

The previous phase successfully diagnosed the hallucination source by isolating the causal anchor $\mathcal{M}_s$. Building upon this, the Intervention phase actively corrects the model's internal representations. We introduce **Bi-Causal Steering**, which constructs counterfactual representations to steer the model's focus toward verified visual evidence across all layers.

3.2.1 Counterfactual Causal Construction

We utilize a pre-trained inpainting model $\mathcal{I}(\cdot,\cdot)$ to construct two complementary counterfactual visual inputs. First, the **Context-Only Image** (I_c) retains only the context: $I_c = \mathcal{I}(I, \mathcal{M}_s)$. Second, the **Anchor-Only Image** (I_a) preserves only the interpretable anchor: $I_a = \mathcal{I}(I, 1 - \mathcal{M}_s)$. Their corresponding features V_c and V_a are extracted via the visual encoder $\mathcal{E}_V$:

$$V_c = \mathcal{E}_V(I_c) \quad ; \quad V_a = \mathcal{E}_V(I_a) \qquad (9)$$

3.2.2 Layer-wise Bi-Causal Steering

We feed V_c and V_a into the VLM decoder. Unlike simple post-hoc interventions, we intervene at every layer $l \in \{1, \ldots, L\}$ to fundamentally rectify the reasoning trajectory. We obtain the layer-wise context-driven states $h_{c,l}^{(t)}$ and anchor-driven states

$h_{a,l}^{(t)}$, where t is introspection conflict step:

$$\begin{aligned} h_{c,l}^{(t)} &= \mathcal{F}_{\text{VLM}}^{(l)}(V_c \oplus \mathcal{E}_T(T_{<t}), h_{c,l-1}^{(t)}) \\ h_{a,l}^{(t)} &= \mathcal{F}_{\text{VLM}}^{(l)}(V_a \oplus \mathcal{E}_T(T_{<t}), h_{a,l-1}^{(t)}) \end{aligned} \tag{10}$$

We define a **correction vector** $\Delta_{h,l}^{(t)}$ for each layer to isolate the pure semantic information contributed by the anchor $\mathcal{M}_s$:

$$\Delta_{h,l}^{(t)} = h_{a,l}^{(t)} - h_{c,l}^{(t)} \tag{11}$$

This vector is injected into the original grounded path at every layer. The debiased state $h_{d,l}^{(t)}$ is computed as:

$$h_{d,l}^{(t)} = h_{g,l}^{(t)} + \alpha \cdot \Delta_{h,l}^{(t)} \tag{12}$$

This multi-layer steering reinforces the model's perception of the key visual region $\mathcal{M}_s$ throughout the entire depth of the network, suppressing biases before they propagate to the final output.

3.2.3 Adaptive Confidence Calibration

We introduce Adaptive Confidence Calibration to mitigate stubborn hallucinations where the model exhibits blind certainty despite lacking distinct visual support. This failure mode is characterized by high global introspection conflict $\mathcal{C}(P_g^{(t)}, P_u^{(t)})$ co-occurring with negligible local divergence between anchor and context states $\mathcal{C}(P_a^{(t)}, P_c^{(t)})$, implying the prediction relies on internal priors rather than the visual anchor.

To suppress this ungrounded confidence, we compute a calibration scalar T_c controlled by a single risk tolerance threshold λ. The penalty activates only when the relative risk exceeds λ, bounded by a hyperbolic tangent to prevent distribution collapse:

$$T_c = 1 + \tanh\left(\max\left(0, \frac{\mathcal{C}(P_g^{(t)}, P_u^{(t)})}{\mathcal{C}(P_a^{(t)}, P_c^{(t)}) + \epsilon} - \lambda\right)\right) \tag{13}$$

where $\epsilon = 10^{-6}$ is a smoothing term. This mechanism adaptively flattens the distribution only when the linguistic prior dominates the visual evidence beyond the allowed tolerance λ. The final corrected probability distribution is obtained by scaling the debiased distribution:

$$P_{\text{corr}}^{(t)} = \sigma(T_c^{-1} \cdot W_o h_{\text{corr},L}^{(t)}), \tag{14}$$

from which the final token is decoded. Theoretical analysis for the invention process is detailed in Appendix A, while latency in Appendix E.

4 Experiments

4.1 Experimental Settings

Benchmarks and Metrics. To comprehensively evaluate the effectiveness of our proposed **VLI** framework, we conduct experiments on two complementary benchmarks: POPE (Li et al., 2023) and MMHal-Bench (Sun et al., 2024). These benchmarks enable a systematic assessment of the model's cognitive introspection capabilities in both discriminative and generative settings.

For POPE, which focuses on object-level discrimination, we follow standard evaluation protocols and report Accuracy and the F_1 score. This benchmark primarily measures the model's ability to correctly identify the presence or absence of objects, providing a fine-grained evaluation of hallucination in discriminative tasks.

However, real-world applications often require free-form text generation rather than binary decisions. To assess hallucination under such scenarios, we further adopt MMHal-Bench, a comprehensive benchmark specifically designed for quantifying the presence and types of hallucinations in complex, open-ended VQA tasks. In MMHal-Bench, model outputs are evaluated by GPT-4 as an automated judge through comparisons with ground-truth object annotations and human-annotated captions. The benchmark reports an overall hallucination score ranging from 0 to 6 following Liu et al. (2025), along with a detailed categorization of different hallucination types (e.g., Attributes, Adversarial, Relations).

Models and Implementation Details. We evaluate the proposed VLI framework on two mainstream large vision–language models, LLaVA-1.5 (Liu et al., 2023) and Qwen3-VL (Bai et al., 2025), using greedy decoding as the default inference strategy. VLI is a training-free, inference-time method that requires no parameter updates. All experiments are implemented in PyTorch. Unless otherwise specified, we follow the same decoding and evaluation protocol for all compared methods. For VLI, we set the anchor energy ratio $\rho = 0.4$ for Attributive Introspection, the introspection conflict threshold θ to 0.1, and the latent steering strength $\alpha = 0.5$ in all experiments, as determined on the validation set.

Baselines. We further compare our approach with advanced baselines spanning three representative hallucination-mitigation paradigms: (i) contrastive decoding methods, including VCD (Leng

Model	Method	MMHAL		POPE (MSCOCO)		POPE (A-OKVQA)		POPE (GQA)	
		Hallu. Rate ↓	Score ↑	Acc (%) ↑	F1 (%) ↑	Acc (%) ↑	F1 (%) ↑	Acc (%) ↑	F1 (%) ↑
LLaVA-1.5	Origin	58.30 ~0.0	2.33 ~0.00	83.82 ~0.00	84.18 ~0.00	79.54 ~0.00	79.81 ~0.00	77.26 ~0.00	77.58 ~0.00
	VCD	63.54 ↑5.24	2.46 ↑0.13	84.67 ↑0.85	85.14 ↑0.96	80.43 ↑0.89	80.92 ↑1.11	78.13 ↑0.87	78.54 ↑0.96
	CICD	58.33 ↑0.03	2.19 ↓0.14	86.46 ↑2.64	87.13 ↑2.95	82.14 ↑2.60	82.93 ↑3.12	79.54 ↑2.28	80.13 ↑2.55
	ClearSight	57.29 ↓1.01	2.16 ↓0.17	88.74 ↑4.92	88.41 ↑4.23	84.58 ↑5.04	84.23 ↑4.42	81.89 ↑4.63	81.64 ↑4.06
	OPERA	58.30 ~0.0	2.40 ↑0.07	88.34 ↑4.52	87.96 ↑3.78	84.13 ↑4.59	83.77 ↑3.96	81.33 ↑4.07	80.97 ↑3.39
	VTI	51.00 ↓7.30	2.39 ↑0.06	87.95 ↑4.13	87.69 ↑3.51	83.87 ↑4.33	83.54 ↑3.73	81.14 ↑3.88	80.88 ↑3.30
	Nullu	54.17 ↓4.13	2.30 ↓0.03	87.18 ↑3.36	86.84 ↑2.66	83.02 ↑3.48	82.54 ↑2.73	80.43 ↑3.17	79.92 ↑2.34
	VLI (Ours)	**45.63** ↓12.67	**3.11** ↑0.78	**89.61** ↑5.79	**89.27** ↑5.09	**85.87** ↑6.33	**85.54** ↑5.73	**83.49** ↑6.23	**83.18** ↑5.60
Qwen3-VL	Origin	40.63 ~0.0	3.56 ~0.00	91.14 ~0.00	90.53 ~0.00	87.13 ~0.00	86.82 ~0.00	82.34 ~0.00	81.93 ~0.00
	VCD	37.50 ↓3.13	3.80 ↑0.24	91.92 ↑0.78	91.37 ↑0.84	87.68 ↑0.55	87.33 ↑0.51	84.84 ↑2.50	84.48 ↑2.55
	CICD	36.46 ↓4.17	3.76 ↑0.20	91.43 ↑0.29	91.08 ↑0.55	87.53 ↑0.40	87.18 ↑0.36	84.13 ↑1.79	83.78 ↑1.85
	ClearSight	39.58 ↓1.05	3.55 ↓0.01	85.04 ↓6.10	83.18 ↓7.35	81.44 ↓5.69	79.82 ↓7.00	79.24 ↓3.10	78.08 ↓3.85
	OPERA	39.10 ↓1.53	3.72 ↑0.16	90.87 ↓0.27	90.28 ↓0.25	86.93 ↓0.20	86.38 ↓0.44	83.54 ↑1.20	83.13 ↑1.20
	VTI	36.46 ↓4.17	3.68 ↑0.12	90.62 ↓0.52	89.94 ↓0.59	86.53 ↓0.60	85.83 ↓0.99	82.68 ↑0.34	81.97 ↑0.04
	Nullu	39.58 ↓1.05	3.53 ↓0.03	88.76 ↓2.38	87.62 ↓2.91	84.93 ↓2.20	83.53 ↓3.29	81.48 ↓0.86	80.93 ↓1.00
	VLI (Ours)	**34.38** ↓6.25	**4.32** ↑0.76	**92.58** ↑1.44	**92.19** ↑1.66	**89.23** ↑2.10	**88.79** ↑1.97	**86.47** ↑4.13	**85.96** ↑4.03

Table 1: Performance evaluation on MMHal and POPE with delta improvements compared to Origin. For POPE, we report the average Accuracy and F1-score across Random, Popular, and Adversarial settings. The best results are highlighted in **bold**.

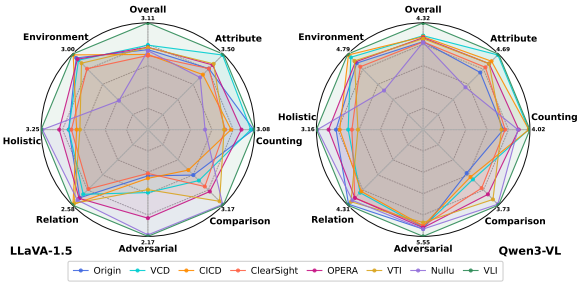

Figure 2: Detailed performance of different models on the eight categories in MMHAL-BENCH, where "Overall" indicates the averaged performance across all categories. A higher score indicates that the generated response contains fewer hallucinations and more information.

et al., 2024) and CICD (Zhao et al., 2025c); (ii) attention-intervention methods, including ClearSight (Yin et al., 2025) and OPERA (Huang et al., 2024); and (iii) latent-space intervention methods, including VTI (Liu et al., 2025) and Nullu (Yang et al., 2025). For all baselines on LLaVA-1.5, we follow the original papers' hyperparameter settings for implementation.

4.2 Main Results

Performance on MMHal-Bench. Table 1 demonstrates that VLI achieves state-of-the-art performance on the challenging open-ended MMHal-Bench. On LLaVA-1.5, VLI reduces the Hallucination Rate (HR) by a substantial **12.67%** (from 58.30% to 45.63%), while on Qwen3-VL, it achieves a record low HR of **34.38%** and the highest overall Score of **4.32**. As illustrated in Fig. 2, VLI yields consistent gains across difficult subsets like *Attribute* and *Adversarial*. These results validate the effectiveness of our *Interpretable Bi-Causal Steering*: unlike decoding methods that passively penalize tokens, our mechanism actively rectifies latent visual-semantic misalignments by isolating specific visual anchors from background noise, which is critical for mitigating fine-grained hallucinations in complex open-ended generation.

Performance on POPE. VLI demonstrates robust generalization across all POPE datasets (MSCOCO, A-OKVQA, and GQA), outperforming baselines in discriminative tasks. On LLaVA-1.5, VLI improves Accuracy by **5.79%** on MSCOCO, and notably achieves even larger gains of **6.33%** and **6.23%** on the more challenging A-OKVQA and GQA datasets, respectively. This superior performance on out-of-distribution and visually complex datasets highlights the advantage of our *Attributive Introspection* mechanism. By precisely localizing causal pixel evidence prior to intervention, VLI avoids the precision-recall trade-off common in global penalty-based decoding (e.g., VCD), allowing it to confidently reject non-existent objects without suppressing valid visual details. Even on the robust Qwen3-VL, VLI further pushes accuracy to **92.58%** on POPE-MSCOCO, proving that introspective grounding remains essential even for stronger base models.

4.3 Ablation Study

To validate the effectiveness of our framework, we conducted an ablation study on MMHal-Bench with LLaVA-1.5 (Table 2). The full VLI yields the

Method	Hallu. Rate ↓	Score ↑
Origin	58.30 ~0.0	2.33 ~0.0
VLI (Ours)	**45.63** ↓12.67	**3.11** ↑0.78
w/o Calibration	47.10 ↓11.20	3.02 ↑0.69
w/o Context only	50.25 ↓8.05	2.82 ↑0.49
w/o Anchor only	53.40 ↓4.90	2.61 ↑0.28

Table 2: Ablation study results on MMHAL-Bench with LLaVA-1.5, with delta improvements compared to Origin.

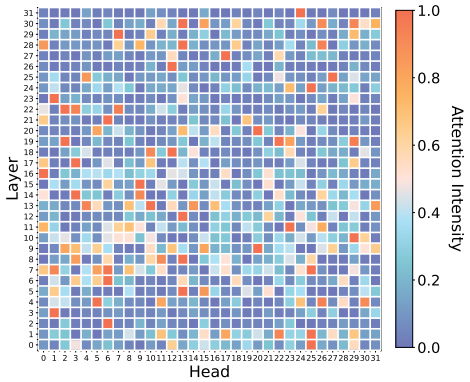

Figure 3: Heatmap of max-pooled attention intensity across layers and heads.

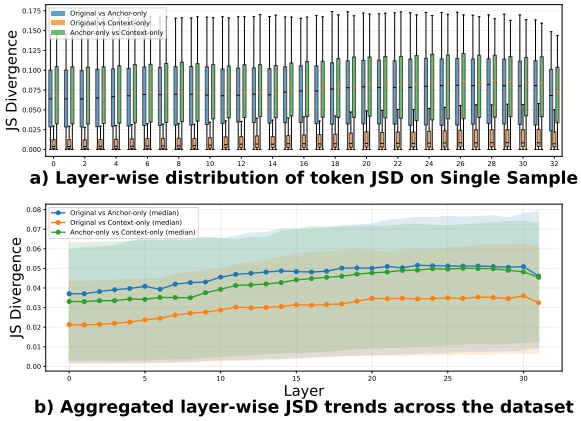

Figure 4: Layer-wise analysis of hidden state shifts. (a) Presenting all tokens in a single representative sample from MMHal-Bench. (b) Focusing on introspected tokens across the MMHal-Bench dataset. The solid lines denote the median JS divergence, while the shaded regions indicate the interquartile range (IQR).

best performance (45.63% hallucination rate, 3.11 score), confirming the synergy of all components.

Dominance of Bi-Causal Steering. The most significant performance degradation occurs in the *w/o Anchor only* setting, where the hallucination rate spikes by 7.77%. This drop outweighs that of the *w/o Context only* variant, indicating that explicitly reinforcing the visual anchor is the primary driver for error rectification. This observation aligns with the layer-wise analysis in Fig. 4, which shows that *Anchor-only* induces a much larger shift in hidden states than *Context-only*. This confirms that the anchor provides the dominant semantic guidance, effectively pulling latent states away from linguistic priors toward the visual ground truth.

Role of Calibration. Conversely, removing *Adaptive Confidence Calibration* results in a relatively minor performance decrease (+1.47%). This is expected, as calibration operates by smoothing the output distribution to penalize ungrounded certainty. Unlike steering, which fundamentally repairs internal semantic representations, calibration serves as a final refinement to prevent the model from being overconfident in its remaining errors.

4.4 Visualization for Introspection and Inventation

Expert Head Attention for Introspection As illustrated in Fig. 3, the attention distribution exhibits a pronounced functional specificity and discreteness. Contrary to a uniform engagement of neural resources, we observe that the vast majority of attention heads remain distinctively silent, represented by cool colors, while high-magnitude activations are concentrated in a sparse subset of layer-specific expert heads. This observation provides strong empirical support for our dynamic head selection mechanism. Since critical semantic information is isolated within these few expert heads, a holistic or average-based approach would inevitably introduce significant noise from the inactive majority. Consequently, dynamically identifying and prioritizing these expert heads is paramount. It allows the model to effectively filter out background interference and establish a precise causal link between the introspection conflict tokens and the relevant visual patches. By focusing solely on these high-activation pathways, our method ensures that the correction process is driven by the most salient visual evidence, rather than dispersed and potentially irrelevant features.

Steering Distance for Inventation To analyze the reliability of our enhanced representations, we compute the token-wise JS divergence between the original hidden states and two counterfactual branches, Anchor-only and Context-only, across all decoder layers (Fig. 4).

In Fig. 4(a), the divergence between Original and Anchor-only states consistently exceeds that

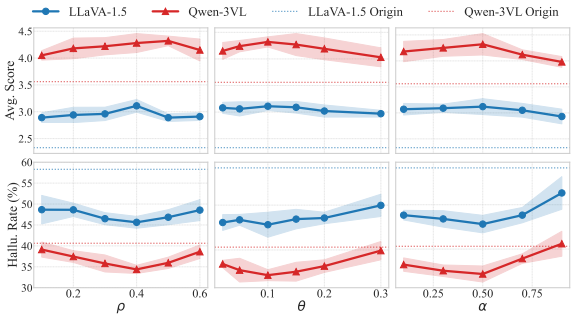

Figure 5: Hyperparameter sensitivity analysis on LLaVA-1.5 and Qwen3-VL. The solid lines represent the average performance of VLI, while the shaded regions indicate the standard deviation, highlighting the stability of our method. The dashed lines denote the baseline performance of the original models (Origin) without intervention.

of the Context-only branch. This implies that removing the anchor alters representations significantly more than removing the context, suggesting the vanilla model is overly shaped by background patterns rather than the critical visual signal—a mechanism consistent with context-driven hallucination. Fig. 4(b) confirms this trend at the dataset level. The median JS divergence for the Anchor-only path increases with depth, significantly surpassing the flat Context-only curve, which indicates an accumulation of context-driven bias. VLI counteracts this drift by explicitly steering hidden states toward the anchor branch, restoring visual grounding. The minimal divergence between original and Context-only states suggests that task-irrelevant background often dominates the global representation. Conversely, the rising deviation of Anchor-only states reveals how accumulated background features progressively displace target semantics. Although this gap narrows in the final layers, the persistent misalignment highlights the necessity of bi-causal steering to reinforce the visual anchor against noise. More analysis for steering JS divergence on all tokens across dataset can be seen in Appendix D.

4.5 Impact of Hyperparameters

We conduct a sensitivity analysis on three key hyperparameters: the cumulative energy ratio ρ for *Attributive Introspection*, the conflict risk threshold θ for triggering intervention, and the steering strength α. The results, illustrated in Fig 5, demonstrate that VLI consistently outperforms the original model across a wide range of settings.

Impact of ρ. The parameter ρ determines the spatial extent of the introspected anchor mask. Performance generally peaks at $\rho = 0.4$, where LLaVA achieves a Score of 3.11 and a hallucination rate of 45.63%. The narrow error bands around the curve suggest that the method remains robust to minor variations in anchor selection. Although Qwen3-VL achieves a slightly higher score at $\rho = 0.5$, its hallucination rate remains lowest at $\rho = 0.4$, confirming the robustness of this setting.

Impact of θ. The risk threshold θ controls the sensitivity of the *Attributive Introspection* phase. Both models achieve optimal performance at $\theta = 0.10$, enabling Qwen3-VL to reach a peak Score of 4.32. Notably, the performance curve remains well above the baseline even at suboptimal thresholds, validating the efficacy of our intervention. However, increasing the threshold beyond 0.15 makes the model too conservative; at $\theta = 0.30$, the LLaVA score drops to 2.97 while the hallucination rate rises to 49.95%, narrowing the performance gap with the baseline.

Impact of α. The steering strength α regulates the magnitude of the latent intervention. The results indicate that a strength of $\alpha = 0.5$ yields the best balance, resulting in a Score of 3.12 for LLaVA and 4.32 for Qwen3-VL. This setting effectively corrects cognitive bias with high stability, as evidenced by the compact error bands. Conversely, excessive steering with $\alpha \geq 0.7$ harms model performance. Specifically, at $\alpha = 0.9$, the LLaVA hallucination rate spikes to 52.65%, causing the performance trajectory to sharply decline towards the baseline level, likely due to over-modification of the hidden states disrupting linguistic fluency.

5 Conclusion

In this paper, we introduced Vision-Language Introspection, a training-free framework designed to mitigate object hallucination by simulating metacognitive self-correction. VLI addresses the disconnect between linguistic priors and visual evidence by synergizing Attributive Introspection for causal anchor localization with Interpretable Bi-Causal Steering for latent representation rectification. This approach effectively isolates visual truths from background noise and neutralizes blind confidence without requiring parameter updates. Experimental results confirm that VLI achieves state-of-the-art performance on both discriminative and generative benchmarks, demonstrating that equipping multimodal models with introspective capabilities offers a robust pathway toward enhanced trustworthiness.

Limitations

First, as analyzed in Appendix E, the construction of counterfactual states during the introspection process introduces additional computational overhead compared to standard decoding strategies. Although we implemented a parallel processing mechanism to mitigate this latency issue, this solution necessitates a significant increase in GPU memory consumption which may constrain deployment on resource-limited devices. Second, the effectiveness of our Attributive Introspection relies on the premise that the base model possesses identifiable expert attention heads that correctly align semantic concepts with visual regions. In scenarios involving highly abstract concepts or where the underlying model fails to form concentrated attention patterns, the precision of causal anchor extraction may degrade and consequently limit the efficacy of the steering intervention.

References

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 2025. 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*, pages 29915–29926.

Kyungho Bae, Jinhyung Kim, Sihaeng Lee, Soonyoung Lee, Gunhee Lee, and Jinwoo Choi. 2025. Mash-vlm: Mitigating action-scene hallucination in video-llms through disentangled spatial-temporal representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13744–13753.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Boxu Chen, Ziwei Zheng, Le Yang, Zeyu Geng, Zhengyu Zhao, Chenhao Lin, and Chao Shen. 2025a. Seeing it or not? interpretable vision-aware latent steering to mitigate object hallucinations. *arXiv preprint arXiv:2505.17812*.

Guoqing Chen, Fu Zhang, Jinghao Lin, Chenglong Lu, and Jingwei Cheng. 2025b. Rrhf-v: Ranking responses to mitigate hallucinations in multimodal large language models with human feedback. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6798–6815.

Jiahe Chen, Jiaying He, Qian Shao, Qiyuan Chen, Jiahe Ying, Hongxia Xu, Jintai Chen, Jianwei Zheng, and Jian Wu. 2025c. Mitigating hallucination of large vision-language models via dynamic logits calibration. *arXiv preprint arXiv:2506.21509*.

Wei Chen, Xin Yan, Bin Wen, Fan Yang, Tingting Gao, Di Zhang, and Long Chen. 2025d. Decoupling contrastive decoding: Robust hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2504.08809*.

Xinlong Chen, Yuanxing Zhang, Qiang Liu, Junfei Wu, Fuzheng Zhang, and Tieniu Tan. 2025e. Mixture of decoding: An attention-inspired adaptive decoding strategy to mitigate hallucinations in large vision-language models. *arXiv preprint arXiv:2505.17061*.

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*.

Jingyuan Deng and Yujiu Yang. 2025. Maskcd: Mitigating lvlm hallucinations by image head masked contrastive decoding. *arXiv preprint arXiv:2510.02790*.

Xinpeng Ding, Kui Zhang, Jianhua Han, Lanqing Hong, Hang Xu, and Xiaomeng Li. 2025. Pami-vdpo: Mitigating video hallucinations by prompt-aware multi-instance video preference learning. *arXiv preprint arXiv:2504.05810*.

Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. 2025. Truthprint: Mitigating lvlm object hallucination via latent truthful-guided pre-intervention. *arXiv preprint arXiv:2503.10602*.

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*, pages 14303–14312.

Yuhan Fu, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Xirong Li. 2025. Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16563–16577.

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. 2025. Cracking the code of hallucination in lvlms with vision-aware head divergence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3488–3501.

Alice Heiman, Xiaoman Zhang, Emma Chen, Sung Eun Kim, and Pranav Rajpurkar. 2025. Factchexcker: Mitigating measurement hallucinations in chest x-ray report generation models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30787–30796.

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*, pages 13418–13427.

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*, pages 27036–27046.

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 Computer Vision and Pattern Recognition Conference*, pages 25004–25014.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, CJ Taylor, and Stefano Soatto. 2024. Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27228–27238.

Jihoon Lee and Min Song. 2025. Retrieval visual contrastive decoding to mitigate object hallucinations in large vision-language models. *arXiv preprint arXiv:2505.20569*.

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*, pages 13872–13882.

Wei Li, Zhen Huang, Houqiang Li, Le Lu, Yang Lu, Xinmei Tian, Xu Shen, and Jieping Ye. 2025a. Visual evidence prompting mitigates hallucinations in large vision-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4048–4080.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N Metaxas. 2025b. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. *arXiv preprint arXiv:2502.03628*.

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

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. 2024a. Phd: A chatgpt-prompted visual hallucination evaluation dataset. *arXiv preprint arXiv:2403.11116*.

Sheng Liu, Haotian Ye, and James Zou. 2025. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024b. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer.

Xinyu Lyu, Beitao Chen, Lianli Gao, Hengtao Shen, and Jingkuan Song. 2024. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *Advances in Neural Information Processing Systems*, 37:122811–122832.

Kyungmin Min, Minbeom Kim, Kang-il Lee, Dongryeol Lee, and Kyomin Jung. 2024. Mitigating hallucinations in large vision-language models via summary-guided decoding. *arXiv preprint arXiv:2410.13321*.

Eunkyu Park, Minyeong Kim, and Gunhee Kim. 2025. Halloc: Token-level localization of hallucinations for vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29893–29903.

Baochang Ren, Shuofei Qiao, Da Zheng, Huajun Chen, and Ningyu Zhang. 2025. Knowrl: Exploring knowledgeable reinforcement learning for factuality. *arXiv preprint arXiv:2506.19807*.

Youxu Shi, Suorong Yang, and Dong Liu. 2025. Exposing hallucinations to suppress them: Vlms representation editing with generative anchors. *arXiv preprint arXiv:2509.21997*.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Li Qing, and Zhaoxiang Zhang. 2025. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12964–12974.

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*, pages 13088–13110.

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. 2025. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29904–29914.

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, et al. 2025. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26147–26159.

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

Tsung-Han Wu, Heekyung Lee, Jiaxin Ge, Joseph E Gonzalez, Trevor Darrell, and David M Chan. 2025a. Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling. *arXiv preprint arXiv:2504.13169*.

Yuanchen Wu, Lu Zhang, Hang Yao, Junlong Du, Ke Yan, Shouhong Ding, Yunsheng Wu, and Xiaoqiang Li. 2025b. Antidote: A unified framework for mitigating lvlm hallucinations in counterfactual presupposition and object perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14646–14656.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. 2025. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25543–25551.

Johnathan Xie, Annie S Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. 2024. Calibrating language models with adaptive temperature scaling. *arXiv preprint arXiv:2409.19817*.

Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. 2024. Mitigating object hallucination via concentric causal attention. *Advances in neural information processing systems*, 37:92012–92035.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. 2025. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14635–14645.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10610–10620.

Zekai Ye, Qiming Li, Xiaocheng Feng, Libo Qin, Yichong Huang, Baohang Li, Kui Jiang, Yang Xiang, Zhirui Zhang, Yunfei Lu, et al. 2025. Claim: Mitigating multilingual object hallucination in large vision-language models with cross-lingual attention intervention. *arXiv preprint arXiv:2506.11073*.

Hao Yin, Guangzong Si, and Zilei Wang. 2025. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14625–14634.

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

Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia Sycara, and Yaqi Xie. 2025. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. *arXiv preprint arXiv:2502.06130*.

Fei Zhao, Chengcui Zhang, Runlin Zhang, Tianyang Wang, and Xi Li. 2025a. Mitigating image captioning hallucinations in vision-language models. *arXiv preprint arXiv:2505.03420*.

Jianfei Zhao, Feng Zhang, Xin Sun, and Chong Feng. 2025b. Aligning attention distribution to information flow for hallucination mitigation in large vision-language models. *arXiv preprint arXiv:2505.14257*.

Jianfei Zhao, Feng Zhang, Xin Sun, Lingxing Kong, Zhixing Tan, and Chong Feng. 2025c. Cross-image contrastive decoding: Precise, lossless suppression of language priors in large vision-language models. *arXiv preprint arXiv:2505.10634*.

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

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

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2025a. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1624–1633.

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

A Theoretical Analysis

In this section, we provide a theoretical foundation for the VLI framework. We demonstrate that our Interpretable Bi-Causal Steering mechanism mathematically functions as a semantic contrastive filter that enhances the Signal-to-Noise Ratio (SNR) of visual representations while decoupling linguistic priors. Furthermore, we justify the Adaptive Confidence Calibration as a regularization of causal sensitivity.

A.1 Latent Linear Representation Hypothesis

Building on recent findings in mechanistic interpretability (Park et al., 2025; Jiang et al., 2025), we posit the *Linear Representation Hypothesis*. We assume that at layer l, the high-dimensional latent state $h \in \mathbb{R}^d$ can be approximated as a linear superposition of independent semantic subspaces. For a multimodal input (V, T), we decompose the latent state h into three orthogonal components:

$$h \approx \mathbf{z}_{obj} + \mathbf{z}_{ctx} + \mathbf{z}_{lang} \tag{15}$$

where:

- $\mathbf{z}_{obj}$: The causal visual vector corresponding to the specific object features defined by the anchor mask $\mathcal{M}_s$.
- $\mathbf{z}_{ctx}$: The visual context vector (background noise/sinks) corresponding to $1 - \mathcal{M}_s$.
- $\mathbf{z}_{lang}$: The linguistic vector encoding syntax and textual priors derived from $T_{<t}$.

Under the hallucination scenario, the model generation y_t is dominated by $\mathbf{z}_{lang}$ and $\mathbf{z}_{ctx}$ (e.g., object co-occurrence priors in the background), while the grounded evidence $\mathbf{z}_{obj}$ is suppressed. Our goal is to rectify the distribution $P(y|h)$ to maximize the mutual information $I(y; \mathbf{z}_{obj})$.

A.2 Derivation of Bi-Causal Steering

VLI constructs two counterfactual states via inpainting: the Context-Only state h_c (where the object is masked) and the Anchor-Only state h_a (where the background is masked).

Assuming the inpainting operation $\mathcal{I}$ effectively suppresses the masked signal to a null vector $\mathbf{0}$ (or a mean vector orthogonal to the specific features), we can formulate these states as:

$$h_g = \mathbf{z}_{obj} + \mathbf{z}_{ctx} + \mathbf{z}_{lang} \tag{16}$$

$$h_c \approx \mathbf{0} + \mathbf{z}_{ctx} + \mathbf{z}_{lang} \tag{17}$$

$$h_a \approx \mathbf{z}_{obj} + \mathbf{0} + \mathbf{z}_{lang} \tag{18}$$

Note that both counterfactual states retain the linguistic component $\mathbf{z}_{lang}$ because the textual input remains identical.

Decoupling Linguistic Priors. We define the steering vector $\Delta = h_a - h_c$. substituting Eq. 17 and 18:

$$\begin{aligned} \Delta &= (\mathbf{z}_{obj} + \mathbf{z}_{lang}) - (\mathbf{z}_{ctx} + \mathbf{z}_{lang}) \\ &= \mathbf{z}_{obj} - \mathbf{z}_{ctx} \end{aligned} \tag{19}$$

Proposition 1 (Linguistic Orthogonality): The steering vector Δ is orthogonal to the linguistic prior $\mathbf{z}_{lang}$. This derivation proves that Δ captures a pure *visual contrast*, the direction pointing towards the object and away from the background, while mathematically canceling out the linguistic priors. This explains why VLI does not degrade language fluency: the steering occurs solely in the visual semantic subspace.

Signal-to-Noise Ratio Enhancement. The rectified state is defined as $h_d = h_g + \alpha\Delta$. Substituting the components:

$$\begin{aligned} h_d &= (\mathbf{z}_{obj} + \mathbf{z}_{ctx} + \mathbf{z}_{lang}) + \alpha(\mathbf{z}_{obj} - \mathbf{z}_{ctx}) \\ &= (1+\alpha)\mathbf{z}_{obj} + (1-\alpha)\mathbf{z}_{ctx} + \mathbf{z}_{lang} \end{aligned} \tag{20}$$

We define the Signal-to-Noise Ratio (SNR) of the visual representation as the ratio of the object magnitude to the context/noise magnitude: $\text{SNR}(h) = \frac{\|\mathbf{z}_{obj}\|}{\|\mathbf{z}_{ctx}\|}$. Comparing the SNR of the grounded state h_g and the rectified state h_d:

$$\text{SNR}(h_d) = \frac{1+\alpha}{1-\alpha} \cdot \frac{\|\mathbf{z}_{obj}\|}{\|\mathbf{z}_{ctx}\|} = \frac{1+\alpha}{1-\alpha} \cdot \text{SNR}(h_g) \tag{21}$$

For any steering strength $0 < \alpha < 1$, the gain factor $\frac{1+\alpha}{1-\alpha} > 1$. **Proposition 2 (SNR Amplification):** The Bi-Causal Steering strictly increases the SNR of the latent state, forcing the model to attend to the causal anchor $\mathbf{z}_{obj}$ while suppressing the confounder $\mathbf{z}_{ctx}$.

A.3 Theoretical Justification for Calibration

The Adaptive Confidence Calibration (Eq. 13) scales the temperature based on the ratio of global conflict to local causal conflict. We formalize this ratio as the *Ungrounded Certainty Ratio*.

Let $\mathcal{D}_{KL}(P||Q)$ denote the divergence. The numerator $\mathcal{C}(P_g, P_u)$ approximates the **Total Perceptual Sensitivity**, how much the model's belief changes given *any* visual input versus no vision. The denominator $\mathcal{C}(P_a, P_c)$ approximates the

Causal Sensitivity, how much the belief changes specifically due to the presence of the object anchor versus the background.

$$R_{risk} = \frac{\text{Total Sensitivity}}{\text{Causal Sensitivity}} \approx \frac{\|\partial P/\partial V\|}{\|\partial P/\partial \mathbf{z}_{obj}\|} \tag{22}$$

Case Analysis:

- **Valid Recognition:** If the model truly sees the object, $\mathcal{C}(P_a, P_c)$ is high (strong causal link). The ratio R_{risk} is low, resulting in $T_c \approx 1$. The distribution remains sharp.
- **Hallucination (Blind Confidence):** If the model predicts an object due to priors or background context, $\mathcal{C}(P_g, P_u)$ may be high (vision changes the prior), but $\mathcal{C}(P_a, P_c) \to 0$ (the specific object pixels do not drive the decision). Here, $R_{risk} \to \infty$.

Consequently, Eq. 13 drives $T_c \gg 1$, maximizing the entropy of the output distribution P_{corr}. This theoretically proves that our calibration mechanism functions as a dynamic regularizer that penalizes predictions unsupported by specific, pixel-wise causal evidence.

B Introspection conflict Analysis

To gain a more concrete understanding of how the introspective conflict score behaves, we analyze nine representative MMHal-Bench examples in Fig. 6, each associated with a specific question. For each example, we compare the grounded decoding path, which conditions on the image, with the ungrounded path, which relies more heavily on language priors. The token-wise JS divergence between these two paths reflects how strongly the model's belief changes once visual evidence is taken into account.

For Case 1 (What color is the fire hydrant cap in the picture?), the question explicitly asks for a color attribute. The conflict curve remains low for function words such as what, color, the, and in, but exhibits a sharp spike on the color token used in the grounded answer (e.g., yellow). This indicates that, without visual grounding, the model tends to follow a strong prior that fire hydrants are typically red, whereas the grounded path adjusts the answer to the correct but less frequent color, producing a high JS divergence exactly at the answer-bearing token. A similar pattern appears in Case 4 (What color are the two cars from right to left in the image?) and Case 6 (What are the colors of the shirts worn by the three men from left to right in the image?), where the largest conflicts occur on the specific color words describing each car or shirt. In Case 8 (What are the colors of the parachutes in the sky?), multiple color tokens show elevated divergence, reflecting that the grounded path must reconcile several distinct colors with the ungrounded prior that tends to favor a small set of frequent colors.

Counting questions show an analogous but complementary behavior. In Case 2 (How many traffic lights are there in the image?), the conflict scores for most tokens are near zero, but the numeral that encodes the predicted count (e.g., four) exhibits a prominent peak. This suggests that language priors alone do not confidently determine the number of traffic lights, and the grounded path must significantly adjust the count based on visual evidence. The same phenomenon is observed in Case 5 (How many bicycles are there in the image?), Case 7 (How many zebras are there in the image?), and Case 9 (How many chairs are there in the image?), where the highest JS divergence consistently concentrates on the numeral token that directly answers the question, while surrounding context tokens remain stable.

Case 3 (How much is it per hour to park at the parking meter?) further illustrates this behavior for fine-grained numeric attributes. Here, the model must output a specific price rather than a small integer count. The conflict curve stays low for the framing tokens (how much, per hour, to park), but spikes on the digits that compose the grounded hourly rate. This indicates that the visual reading of the meter substantially revises the ungrounded guess about the price, again localizing conflict to the answer-bearing portion of the sequence.

Across all nine cases, we observe a consistent sparsity pattern: most tokens have JS divergence well below the threshold $\theta = 0.10$, and only a small number of semantically critical tokens, colors, numerals, or digits that directly respond to the question, exceed this cutoff. This supports our design of the introspective conflict score as a selective trigger rather than a global perturbation: it remains quiet on benign context tokens and becomes active only where vision–language mismatch is likely to cause hallucination. Combined with our earlier hyperparameter study, which shows that thresholds in the vicinity of $\theta = 0.1$ yield the best trade-off between hallucination reduction and overall per-

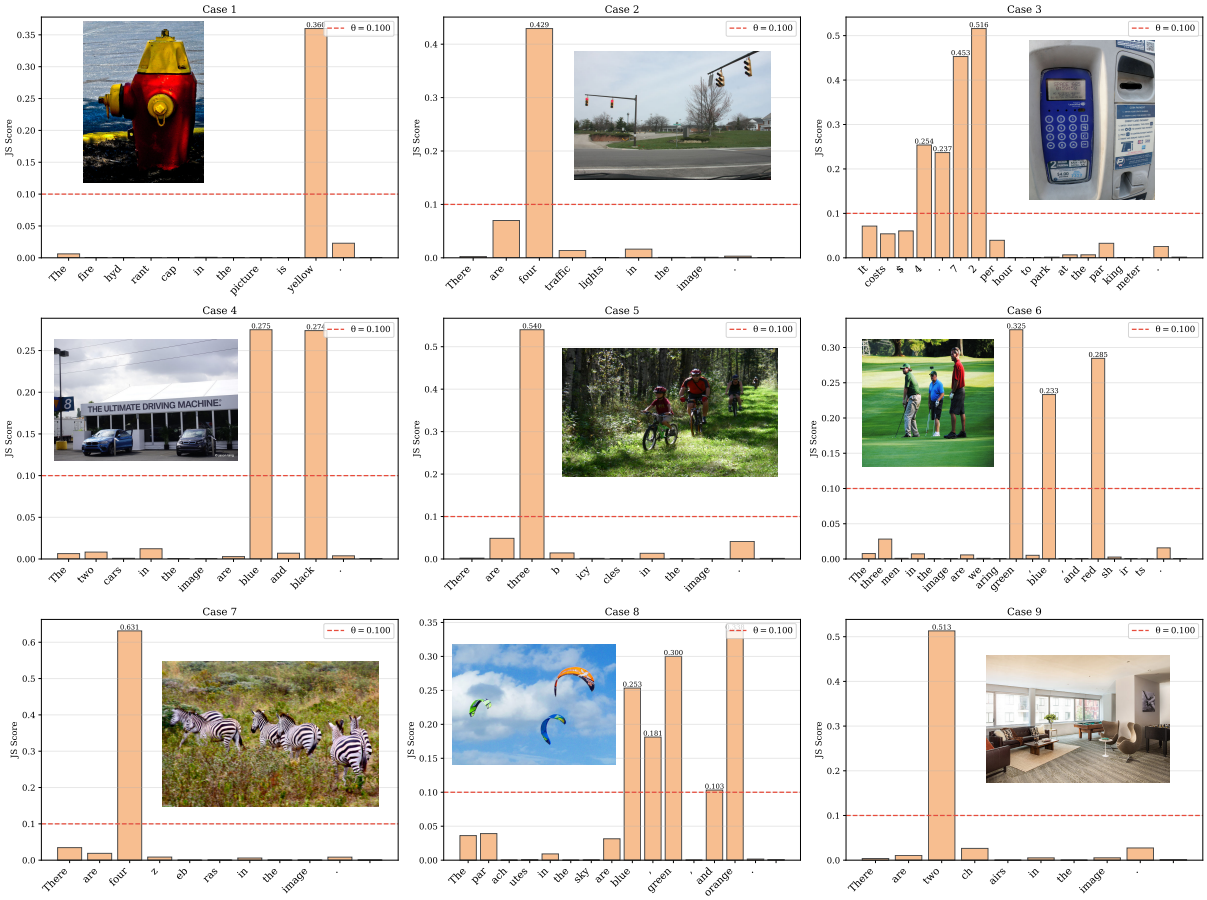

Figure 6: Introspective conflict scores between grounded and ungrounded decoding paths. Token-wise introspective conflict scores of LLaVA-1.5 on nine representative MMHal-Bench samples. Each panel shows the input image together with the grounded answer, where the height of each bar indicates the JS divergence of the corresponding token between the grounded and ungrounded paths. The red dashed line marks the conflict-risk threshold $\theta = 0.10$ used to trigger introspection.

formance, these case studies provide qualitative evidence that the chosen threshold is both effective and reasonable. It allows VLI to focus introspective interventions precisely on the answer tokens where correcting hallucinations matters most.

C Robustness to Visual Attention Sinks

Recent studies (Kang et al., 2025) identify the visual attention sink phenomenon in LVLMs, where specific irrelevant tokens such as delimiters or background patches disproportionately absorb attention mass. A standard identification method relies on detecting anomalous activation magnitudes:

$$\phi(v_j) = \max_{d \in \mathcal{D}_{sink}} \frac{|v_j[d]|}{\|v_j\|_2} \tag{23}$$

where tokens exceeding a threshold τ_{sink} are flagged as sinks.

While explicit masking of these sinks is a common remedy, our **Expert Head Selection** and **Cumulative Energy Thresholding** mechanisms described in Sec 3.1.3 provide intrinsic robustness against this noise without requiring a separate sink detection module. Since visual sinks typically do not align with the semantic regions required for grounded prediction, they are naturally filtered out during the expert head calibration phase characterized by low $\mu_{l,h}$ scores. Furthermore, our adaptive anchor extraction focuses on the cumulative probability mass ρ. Consequently, unless a sink dominates the global attention distribution to an extreme degree, which is a rare occurrence in the identified expert heads, it is excluded from the causal mask $\mathcal{M}_s$. Therefore, VLI efficiently purifies visual evidence without the computational overhead of explicit sink modeling.

To validate this, we compare our standard VLI framework against a variant augmented with explicit sink masking, denoted as *VLI + Explicit Sink Masking*, and two ablated variants lacking our core filtering mechanisms. The results are reported in Table 3.

Method Configuration	MMHal-Bench		POPE (MSCOCO)		POPE (A-OKVQA)		POPE (GQA)	
	Score ↑	Hallu. Rate ↓	Acc (%) ↑	F1 (%) ↑	Acc (%) ↑	F1 (%) ↑	Acc (%) ↑	F1 (%) ↑
Base Model: LLaVA-1.5								
Origin (Baseline)	2.33	58.30	83.82	84.18	79.81	79.54	77.58	77.26
VLI (Standard)	**3.11**	**45.63**	**89.27**	**89.61**	**85.87**	**85.54**	**83.18**	**83.49**
VLI + Explicit Sink Masking	3.12	45.58	89.29	89.65	85.91	85.60	83.22	83.55
VLI w/o Expert Heads (Avg)	2.65	52.14	86.05	86.44	82.10	81.85	80.05	79.90
VLI w/o Adaptive Threshold (Fixed-k)	2.89	48.75	87.55	87.90	83.45	83.10	81.30	81.15
Base Model: Qwen3-VL								
Origin (Baseline)	3.56	40.63	90.53	91.14	86.82	87.13	82.34	81.93
VLI (Standard)	**4.32**	**34.38**	**92.58**	**92.19**	**88.79**	**89.23**	**85.96**	**86.47**
VLI + Explicit Sink Masking	4.33	34.32	92.61	92.25	88.85	89.30	86.01	86.52
VLI w/o Expert Heads (Avg)	3.88	37.95	91.20	91.55	87.40	87.65	83.50	83.80
VLI w/o Adaptive Threshold (Fixed-k)	4.10	36.12	91.85	91.80	88.05	88.45	84.80	85.10

Table 3: Robustness analysis against Visual Attention Sinks on MMHal-Bench and POPE. We compare our standard VLI against variants with explicit sink masking and ablated attention mechanisms. **VLI + Sink Masking** explicitly filters tokens based on Eq. 23. The negligible performance gap validates that VLI is intrinsically robust to attention sinks.

Negligible Gain from Explicit Masking. Comparing *VLI (Standard)* with *VLI + Explicit Sink Masking*, we observe minimal performance differences across all metrics. For instance, on LLaVA-1.5, the hallucination rate on MMHal-Bench improves marginally from 45.63% to 45.58% where $\Delta < 0.1\%$, and POPE accuracy remains statistically stagnant. This confirms that the tokens identified as sinks by explicit algorithms are already being effectively filtered out by the internal mechanisms of VLI, rendering the additional computational overhead of sink detection redundant.

Role of Expert Head Selection. The significant performance drop in *VLI w/o Expert Heads*, exemplified by a 6.51% increase in hallucination rate on LLaVA-1.5, highlights the critical role of our head selection strategy. Attention sinks typically manifest as high-magnitude activations across global average heads. By selectively aggregating attention only from *expert heads* $\mathcal{H}_{expert}$ that demonstrate high localization accuracy $\mu_{l,h}$, VLI naturally bypasses heads dominated by sink tokens, thereby purifying the causal signal.

Efficacy of Cumulative Energy Thresholding. Similarly, replacing our adaptive energy thresholding with a fixed top-k strategy, referred to as *VLI w/o Adaptive Threshold*, leads to a noticeable performance degradation. Fixed-k selection risks including high-activation sink tokens that may appear in the long tail of the distribution, or excluding valid semantic regions when the object is large. Our cumulative energy approach utilizing ρ ensures that the anchor mask $\mathcal{M}_s$ locks onto the semantic core, naturally excluding sink tokens unless they dominate the probability mass. Such domination is a rarity within identified expert heads.

In conclusion, VLI achieves robustness to visual attention sinks not through external patching, but through the synergistic design of expert head selection and adaptive anchor extraction.

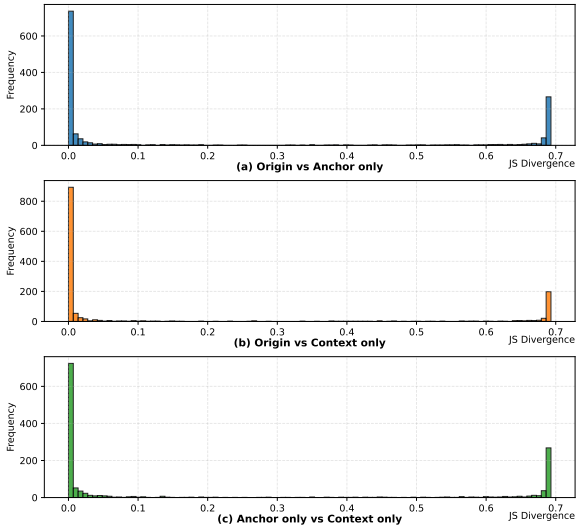

Figure 7: Distribution of Jensen–Shannon divergence between the logits produced by the original model and those produced using the enhanced input across all samples.

D Logits Divergence Analysis

Fig. 7 presents the histograms of Jensen-Shannon (JS) divergence between the logits of the original decoding path and the counterfactual steering branches. The distribution exhibits a pronounced bimodality characterized by a heavy concentration of mass near zero and a distinct, sparse peak in the high-divergence region around 0.7. This statistical behavior provides empirical validation for the three core components of the VLI framework:

Attributive Introspection, Bi-Causal Steering, and Adaptive Confidence Calibration.

First, the overwhelming density near zero divergence in Fig. 7(a) and Fig. 7(b) indicates that for the vast majority of generated tokens, the linguistic priors and visual context remain consistent. These tokens correspond to functional words or unambiguous context, as qualitatively illustrated in Fig. 6, where tokens such as *there*, *are*, and *in* exhibit negligible conflict scores. This functional sparsity confirms that computationally expensive interventions are unnecessary for most generation steps, justifying the efficiency of our selective introspection mechanism which only triggers upon detecting conflict.

Second, the secondary peak in the high-divergence region signifies the existence of a specific subset of tokens where the visual anchor strictly contradicts the background context and linguistic priors. This aligns with the findings in Fig. 4(b), where the Anchor-only hidden states progressively deviate from the Context-only states across model layers. The sharp separation in Fig. 7(c) between Anchor-only and Context-only logits proves that hallucinations are not caused by global degradation but by specific conflicts where the background noise overwhelms the object signal. This precise separation is critical for the efficacy of our Bi-Causal Steering vector Δh, ensuring that the intervention vector is orthogonal to the linguistic subspace and targets only the semantic misalignment.

Finally, the clear bimodality supports the design of the Adaptive Confidence Calibration mechanism. The discrete nature of the high-divergence mode suggests that hallucination is a binary state change rather than a linear degradation. Consequently, the use of the hyperbolic tangent function in Eq. 13 is theoretically sound, as it functions as a soft-gate that rapidly penalizes confidence only when the token falls into this high-risk distribution tail. The ablation study in Table 2 further corroborates this; the removal of the Anchor-only component results in the most significant performance drop because it eliminates the high-divergence signal necessary to counteract the context-driven hallucinations prevalent in the original distribution.

E Latency and Computational Cost Analysis

Table 4 presents a comprehensive latency comparison across different hallucination mitigation paradigms. Our proposed VLI framework, when optimized, achieves a competitive latency profile that balances computational efficiency with the depth of cognitive introspection.

Serial vs. Parallel Inference. As indicated in Table 4, the naive serial implementation of VLI (*VLI Serial*) results in an average per-token latency of 216.22 ms. This increased latency is inherent to the *Bi-Causal Steering* mechanism, which requires the computation of two additional counterfactual distinct states—Anchor-Only (h_a) and Context-Only (h_c)—alongside the original decoding path to derive the steering vector Δh. In a serial execution regime, these forward passes are computed sequentially, effectively tripling the inference cost for every generated token where introspection is triggered.

To mitigate this bottleneck, we implement a parallel processing mechanism (*VLI Parallel*) that reduces the average per-token latency to 95.41 ms. This represents a speedup of approximately $2.27\times$ compared to the serial version, bringing our method's latency close to that of standard contrastive decoding methods like VCD (76.28 ms) and significantly lower than heavy attention-intervention methods such as OPERA (405.56 ms) or ClearSight (826.49 ms).

Parallelization Mechanism. The parallelization is achieved by exploiting the independence of the counterfactual branches. During the *Interpretable Bi-Causal Steering* phase, the computation of the anchor-driven state h_a and the context-driven state h_c does not depend on their mutual intermediate outcomes within the same step. Consequently, we construct a consolidated batch input $V_{batch} = [V_{original}; V_{anchor}; V_{context}]$ effectively performing the forward passes for all three representations simultaneously within a single GPU operation. This allows VLI to leverage the massive parallelism of modern hardware, masking the latency overhead of the additional forward passes.

Memory Overhead and Trade-offs. While parallelization significantly reduces inference time, it introduces a trade-off regarding memory consumption. By processing the original and counterfac-

Method	Efficiency Metrics						Performance (MMHal)	
	Total Runtime (s)	Total Tokens	Latency (ms/token) ↓	Relative Cost ↓	Throughput (tokens/s) ↑	Memory Overhead	Hallucination Rate (%) ↓	Overall Score ↑
Origin (Baseline)	3.130	70	**44.71**	**1.00**×	**22.37**	1.0×	58.30	2.33
VCD	6.179	81	76.28	1.71×	13.11	∼2.0×	63.54	2.46
CICD	9.308	80	116.35	2.60×	8.59	∼2.0×	58.33	2.19
ClearSight	66.119	80	826.49	18.49×	1.21	∼1.0×	57.29	2.16
OPERA	30.011	74	405.56	9.07×	2.47	∼1.0×	58.30	2.40
VTI	6.853	80	85.66	1.92×	11.67	1.0×	51.00	2.39
Nullu	9.804	74	132.48	2.96×	7.55	1.0×	54.17	2.30
VLI (Serial)	17.730	82	216.22	4.84×	4.62	1.0×	**45.63**	**3.11**
VLI (Parallel)	7.823	82	95.41	2.13×	10.48	∼3.0×	**45.63**	**3.11**

Table 4: **Comprehensive Efficiency and Performance Analysis.** This table integrates raw runtime data with derived efficiency metrics and performance outcomes. *Relative Cost* denotes the latency multiplier relative to the Origin model. *Memory Overhead* is estimated based on the requirement for parallel decoding streams (e.g., VLI Parallel processes anchor, context, and grounded states simultaneously).

tual streams in a single batch, the peak memory footprint increases during the steering phase. This overhead stems primarily from two sources:

1. **Activation Storage:** The model must store intermediate activations for three concurrent streams instead of one, essentially tripling the memory required for temporary tensors during the forward pass.

2. **KV-Cache Expansion:** To maintain context for the counterfactual paths across generation steps, the Key-Value (KV) cache must be maintained for the h_a and h_c branches in addition to the h_g branch. This results in a linear increase in video memory usage proportional to the number of parallel streams.

Despite this increased memory demand, the parallelized VLI framework remains deployable on standard academic hardware setups. The analysis confirms that by accepting a manageable increase in memory occupancy, VLI achieves a favorable sweet spot: it **delivers state-of-the-art hallucination reduction with a latency cost only marginally higher than simple contrastive baselines**, avoiding the prohibitive slowness of iterative attention-editing approaches.

F Case Study

F.1 Interpretable Anchors in Attributive Introspection Phase

In the *Attributive Introspection* phase, the primary objective of VLI is to trace abstract cognitive dissonance, where linguistic priors conflict with sensory inputs, back to concrete regions in the input image. This process transforms latent cognitive uncertainty into explicit, interpretable visual evidence. This introspection culminates in the construction of a pixel-precise **Causal Anchor Mask**, which provides a grounded basis for the subsequent bi-causal steering.

The core intermediate representation in this phase is the Purified Attention Map, produced by the Attention Purification module. This module is designed to eliminate systemic and structural biases in raw attention by combining Expert Head Selection with Visual Sink Suppression. The resulting purified attention more faithfully reflects genuine visual grounding relevant to the token triggering the conflict.

Based on this purified signal, VLI generates the Causal Anchor Mask via a Cumulative Energy Thresholding strategy. Rather than relying on fixed thresholds, this adaptive mechanism selects the minimal set of pixels whose cumulative attention energy explains the dominant semantics of the introspection target, while effectively suppressing long-tail noise. The resulting binary mask thus captures only the visually critical regions responsible for verifying the model's prediction.

As illustrated in Fig. 8, VLI consistently localizes interpretable anchors across diverse visual reasoning tasks. In counting tasks (Case 3: determining the number of chairs), the framework introspects and highlights distinct object contours corresponding to each counted instance. For attribute recognition tasks (Case 2: identifying parachute colors and Case 5: identifying the color of the left ball), VLI isolates multiple spatially dispersed semantic targets across a wide field of view. In

fine-grained detail recognition scenarios (Case 1: identifying which cat has its mouth open and Case 4: identifying who is wearing trousers), VLI precisely focuses on the relevant anatomical region rather than the entire object, demonstrating high-resolution semantic alignment. These results validate that VLI moves beyond black-box failure analysis by establishing an explicit correlation between high-conflict tokens and their corresponding visual stimuli.

By explicitly mapping internal cognitive conflicts to external visual regions, VLI moves beyond treating hallucinations as black-box failures. The resulting anchor masks provide direct visual evidence that the identified tokens are intrinsically correlated with specific visual anchors in the input, validating the introspection process.

F.2 Interpretable Bi-Causal Steering in Intervention Phase

We provide additional examples through Fig. 9 to Fig. 13 showing the effect of the *Interpretable Bi-Causal Steering* phase by detailing the evolution of logits during the generation process. These examples illustrate that VLI exerts a substantial influence on the final logit distribution by dynamically contrasting evidence against background noise. By effectively rectifying the probability bias at critical decision steps, our approach successfully steers the model away from overconfident linguistic priors and ensures the output aligns with the visual facts.

In Fig. 9, despite the clear dining setting, the baseline model hallucinates the presence of sand, initiating its response with "Yes" (0.6518). VLI effectively reverses this error at the logit level. At Step 1, it suppresses the "Yes" token to 0.4073 and elevates the correct "No" token to 0.5655. Furthermore, at Step 5, while the baseline model remains ambiguous, predicting "a" (0.6183) or "sand" (0.3578), our method solidly predicts "no" with a near-certain probability of 0.9986, ensuring the generated answer accurately states "there is no sand".

Fig. 10 involves an image of surfers standing on a sandy beach, prompting the question, "Is there a grass in the image?". The baseline model misinterprets the texture of the ground, leading to an object hallucination where it asserts, "Yes, there is a grassy area". The logits table reveals that the baseline model strongly commits to this error at Step 1, assigning a probability of 0.6177 to the token "Yes". VLI successfully rectifies this at the onset by suppressing the affirmative response and elevating the correct token "No" to the top rank with a probability of 0.5260. The intervention remains robust at Step 5, where our method assigns a near-certain probability of 0.9979 to the token "no" (completing the phrase "there is no grass"), whereas the baseline remains confused, splitting its probability between "a" (0.4764) and "grass" (0.4337).

Visual ambiguity can often trigger hallucinations, as demonstrated by Fig. 11 of a train-shaped cake, where the supporting table is mislabeled as a "cabinet". Unlike the previous clear-cut errors, the baseline here is initially ambivalent, only marginally preferring the incorrect "Yes" (0.5066) over "No" (0.4613) at Step 1. VLI proves decisive in these borderline cases. By effectively inverting the probability distribution via anchor reinforcement, it secures the correct trajectory with a "No" prediction (0.5397). This early intervention prevents the cascade of errors seen in the baseline, which at Step 5 firmly commits to the hallucination (predicting "a" with 0.9689 probability). In contrast, our method solidifies the correction with a definitive "no" (0.9962), ensuring the generated caption accurately reflects the scene.

Fig. 12 illustrates a scenario where the model misinterprets the geometry of the scene, confusing a steep snowy slope for a vertical structure. When asked, "Is there a wall in the image?", the baseline model incorrectly affirms the presence of a wall, likely conflating the solid white expanse of the hill with a built barrier. The logit analysis at Step 1 shows the baseline favoring the hallucination with a "Yes" probability of 0.5390, compared to 0.4331 for "No". VLI effectively intervenes to correct this spatial misunderstanding. It shifts the probability distribution to favor "No" (0.6036), suppressing the "Yes" token to 0.3605. This correction prevents the model from constructing the erroneous phrase "a wall"; instead, at Step 5, our method assigns a decisive 0.9969 probability to the token "no", ensuring the final output correctly reflects the absence of the object.

Fig. 13 presents a challenging case involving a skier lying on the snow, where the background contains wooden fencing and barriers that act as visual confounders. The baseline model is misled by these wooden textures, likely misidentifying the wooden slats as parts of a chair, and confidently pre-

dicts "Yes" (0.5894) at Step 1. VLI demonstrates its specific capability to target and resolve such ambiguity. By identifying the causal visual regions responsible for this confusion and enhancing the semantic contrast, VLI effectively suppresses the activation of erroneous features associated with the wooden elements. Consequently, the logit distribution at Step 1 is corrected to favor "No" (0.5374). This targeted intervention ensures that the model distinguishes the background noise from the foreground subject, resulting in a firm "no" prediction (0.9979) at Step 5 and a correct description of the person lying on the ground.

G Usage of AI Assistant

The authors acknowledge the use of Gemini 3 Pro to assist with language editing and grammatical corrections. We affirm that the AI tool was not involved in the generation of scientific ideas, formulation of the methodology, or interpretation of the data. All intellectual content remains the work of the human authors.

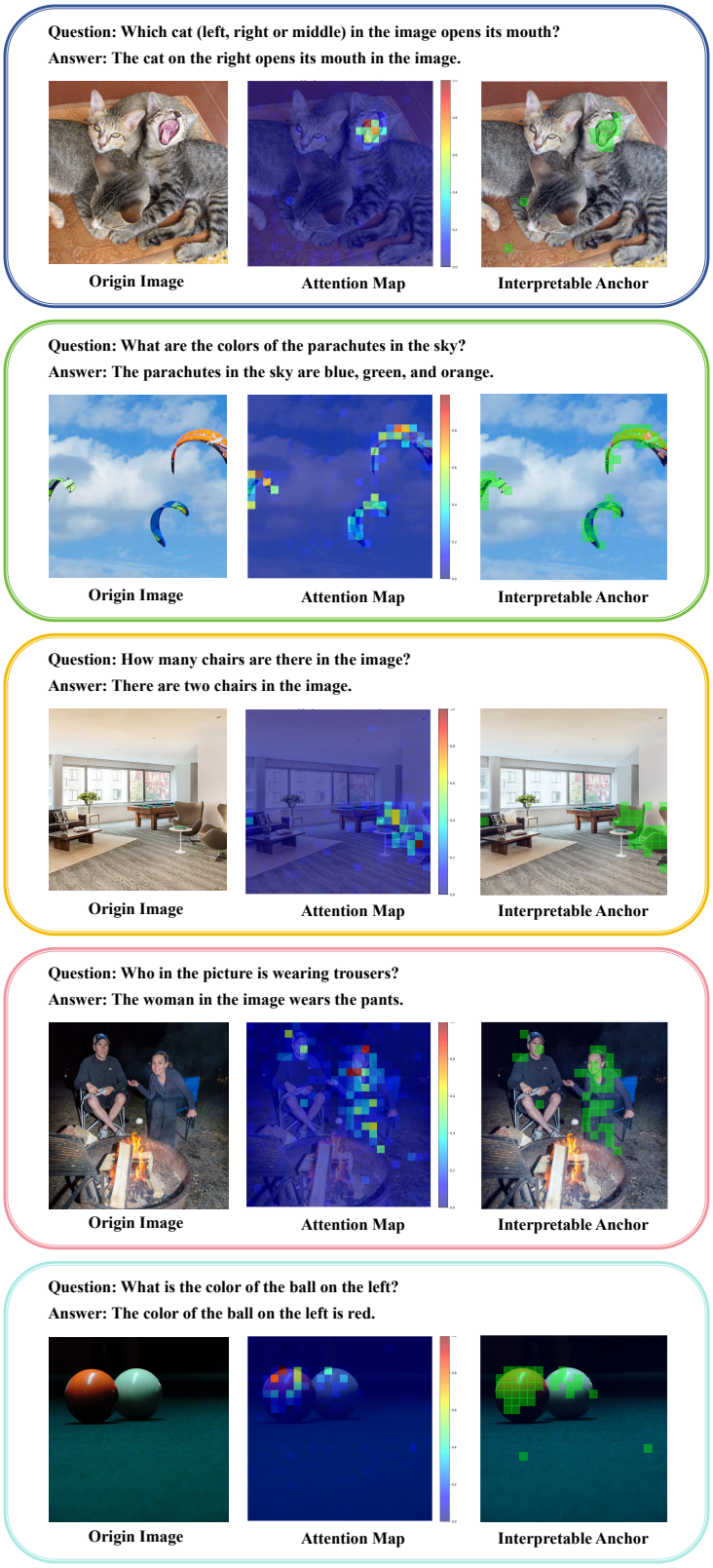

Figure 8: Visual comparisons of introspected visual anchor patterns under different scenarios.

Question: Is there a sand in the image?

Ground Truth: No

Original Answer:
Yes, there is a sand in the image, and the two children are sitting at a table with plates of food and drinks.

Corrected Answer:
No, there is no sand in the image. The image features a young boy and a girl sitting at a table with plates of food, glass.

Logits Changes (29 Generation Steps)

Step 1

B	Token	Prob	A	Token	Prob
1	Yes	0.6518	1	No	0.5655
2	No	0.3226	2	Yes	0.4073
3	There	0.0166	3	There	0.0179
4	The	0.0047	4	The	0.0050
5	While	0.0011	5	While	0.0010

Step 2

B	Token	Prob	A	Token	Prob
1	,	0.9999	1	,	0.9995
2	there	0.0000	2	.	0.0003
3	.	0.0000	3	there	0.0002
4	—	0.0000	4	sand	0.0001
5	!	0.0000	5	—	0.0000

Step 3

B	Token	Prob	A	Token	Prob
1	there	0.9213	1	there	0.9742
2	the	0.0445	2	the	0.0240
3	a	0.0154	3	sand	0.0006
4	in	0.0123	4	I	0.0003
5	sand	0.0020	5	it	0.0002

Step 4

B	Token	Prob	A	Token	Prob
1	is	0.9513	1	is	0.9418
2	are	0.0459	2	isn	0.0440
3	’	0.0024	3	are	0.0111
4	appears	0.0002	4	’	0.0028
5	seems	0.0000	5	aren	0.0001

Step 5

B	Token	Prob	A	Token	Prob
1	a	0.6183	1	no	0.9986
2	sand	0.3578	2	not	0.0010
3	some	0.0083	3	neither	0.0001
4	an	0.0074	4	none	0.0001
5	play	0.0005	5	a	0.0001

Step 6

B	Token	Prob	A	Token	Prob
1	sand	0.8776	1	sand	0.9977
2	p	0.0208	2	beach	0.0005
3	small	0.0197	3	visible	0.0003
4	little	0.0149	4	image	0.0002
5	young	0.0062	5	snow	0.0001

Figure 9: Case 1 from the POPE Random split (GQA subset).

Question: Is there a grass in the image?

Ground Truth: No

Original Answer:
Yes, there is a grassy area in the image, where the group of people is standing.

Corrected Answer:
No, there is no grass in the image. It is a beach scene with people and surfboards.

Logits Changes (21 Generation Steps)

Step 1

B	Token	Prob	A	Token	Prob
1	Yes	0.6177	1	No	0.5260
2	No	0.3465	2	Yes	0.4227
3	There	0.0236	3	There	0.0336
4	The	0.0079	4	The	0.0116
5	While	0.0010	5	While	0.0012

Step 2

B	Token	Prob	A	Token	Prob
1	,	0.9999	1	,	0.9996
2	there	0.0000	2	.	0.0003
3	.	0.0000	3	there	0.0001
4	—	0.0000	4	—	0.0000
5	;	0.0000	5	grass	0.0000

Step 3

B	Token	Prob	A	Token	Prob
1	there	0.9551	1	there	0.9354
2	the	0.0359	2	the	0.0637
3	in	0.0046	3	in	0.0001
4	a	0.0015	4	grass	0.0001
5	grass	0.0009	5	it	0.0001

Step 4

B	Token	Prob	A	Token	Prob
1	is	0.9505	1	is	0.9457
2	are	0.0466	2	isn	0.0372
3	’	0.0024	3	are	0.0144
4	appears	0.0004	4	’	0.0024
5	seems	0.0001	5	aren	0.0001

Step 5

B	Token	Prob	A	Token	Prob
1	a	0.4764	1	no	0.9979
2	grass	0.4337	2	not	0.0011
3	some	0.0487	3	none	0.0004
4	green	0.0142	4	grass	0.0002
5	an	0.0070	5	a	0.0002

Step 6

B	Token	Prob	A	Token	Prob
1	grass	0.6617	1	grass	0.9932
2	patch	0.1454	2	visible	0.0047
3	small	0.0552	3	green	0.0013
4	green	0.0265	4	patch	0.0001
5	field	0.0168	5	l	0.0001

Figure 10: Case 2 from the POPE adversarial split (GQA subset).

Question: Is there a cabinet in the image?

Ground Truth: No

Original Answer:
Yes, there is a cabinet in the image, and the cake is placed on top of it.

Corrected Answer:
No, there is no cabinet in the image. The image features a cake shaped like a train on a table.

Logits Changes (22 Generation Steps)

Step 1

B	Token	Prob	A	Token	Prob
1	Yes	0.5066	1	No	0.5397
2	No	0.4613	2	Yes	0.4269
3	There	0.0256	3	There	0.0277
4	The	0.0041	4	The	0.0036
5	While	0.0007	5	While	0.0005

Step 2

B	Token	Prob	A	Token	Prob
1	,	0.9999	1	,	0.9996
2	.	0.0000	2	.	0.0003
3	there	0.0000	3	there	0.0001
4	—	0.0000	4	—	0.0000
5	;	0.0000	5	;	0.0000

Step 3

B	Token	Prob	A	Token	Prob
1	there	0.9358	1	there	0.9842
2	the	0.0431	2	the	0.0152
3	a	0.0144	3	I	0.0002
4	in	0.0053	4	it	0.0001
5	an	0.0003	5	this	0.0000

Step 4

B	Token	Prob	A	Token	Prob
1	is	0.9970	1	is	0.9202
2	are	0.0020	2	isn	0.0755
3	’	0.0007	3	are	0.0031
4	appears	0.0002	4	’	0.0010
5	seems	0.0000	5	aren	0.0000

Step 5

B	Token	Prob	A	Token	Prob
1	a	0.9689	1	no	0.9962
2	an	0.0288	2	not	0.0035
3	at	0.0003	3	neither	0.0001
4	cabinet	0.0003	4	none	0.0001
5	one	0.0002	5	a	0.0000

Step 6

B	Token	Prob	A	Token	Prob
1	cabinet	0.7543	1	cabinet	0.9978
2	glass	0.1070	2	visible	0.0008
3	white	0.0254	3	cabin	0.0005
4	wooden	0.0124	4	cup	0.0001
5	large	0.0113	5	wooden	0.0001

Figure 11: Case 3 from the POPE adversarial split (GQA subset).

Question: Is there a wall in the image?

Ground Truth: No

Original Answer:
Yes, there is a wall in the image, and the person is skiing down it.

Corrected Answer:
No, there is no wall in the image. The image features a person skiing down a snowy hill.

Logits Changes (20 Generation Steps)

Step 1

B	Token	Prob	A	Token	Prob
1	Yes	0.5390	1	No	0.6036
2	No	0.4331	2	Yes	0.3605
3	There	0.0206	3	There	0.0278
4	The	0.0043	4	The	0.0048
5	In	0.0010	5	While	0.0008

Step 2

B	Token	Prob	A	Token	Prob
1	,	0.9999	1	,	0.9995
2	.	0.0000	2	.	0.0003
3	there	0.0000	3	there	0.0001
4	—	0.0000	4	—	0.0001
5	;	0.0000	5	wall	0.0000

Step 3

B	Token	Prob	A	Token	Prob
1	there	0.9173	1	there	0.9825
2	the	0.0654	2	the	0.0169
3	in	0.0102	3	I	0.0002
4	a	0.0060	4	in	0.0001
5	an	0.0002	5	it	0.0001

Step 4

B	Token	Prob	A	Token	Prob
1	is	0.9982	1	is	0.9399
2	’	0.0010	2	isn	0.0573
3	are	0.0006	3	’	0.0014
4	appears	0.0001	4	are	0.0012
5	seems	0.0000	5	doesn	0.0000

Step 5

B	Token	Prob	A	Token	Prob
1	a	0.9824	1	no	0.9969
2	an	0.0152	2	not	0.0028
3	snow	0.0013	3	neither	0.0001
4	wall	0.0001	4	none	0.0000
5	white	0.0001	5	a	0.0000

Step 6

B	Token	Prob	A	Token	Prob
1	wall	0.4967	1	wall	0.9964
2	snow	0.1117	2	visible	0.0022
3	large	0.0898	3	physical	0.0003
4	mountain	0.0421	4	actual	0.0002
5	white	0.0414	5	solid	0.0001

Figure 12: Case 4 from the POPE adversarial split (GQA subset).

Question: Is there a chair in the image?

Ground Truth: No

Original Answer:
Yes, there is a chair in the image, and a person is sitting on it.

Corrected Answer:
No, there is no chair in the image. The image shows a person lying on the ground in the snow.

Logits Changes (19 Generation Steps)

Step 1

B	Token	Prob	A	Token	Prob
1	Yes	0.5894	1	No	0.5374
2	No	0.3746	2	Yes	0.4120
3	There	0.0308	3	There	0.0441
4	The	0.0027	4	The	0.0039
5	In	0.0009	5	While	0.0006

Step 2

B	Token	Prob	A	Token	Prob
1	,	1.0000	1	,	0.9996
2	there	0.0000	2	.	0.0002
3	.	0.0000	3	there	0.0001
4	—	0.0000	4	—	0.0000
5	;	0.0000	5	;	0.0000

Step 3

B	Token	Prob	A	Token	Prob
1	there	0.9730	1	there	0.9932
2	the	0.0091	2	the	0.0065
3	a	0.0077	3	I	0.0001
4	in	0.0077	4	it	0.0000
5	someone	0.0007	5	in	0.0000

Step 4

B	Token	Prob	A	Token	Prob
1	is	0.9936	1	is	0.9364
2	are	0.0051	2	isn	0.0562
3	’	0.0012	3	are	0.0063
4	appears	0.0001	4	’	0.0010
5	seems	0.0000	5	aren	0.0000

Step 5

B	Token	Prob	A	Token	Prob
1	a	0.9520	1	no	0.9979
2	an	0.0438	2	not	0.0019
3	someone	0.0018	3	neither	0.0001
4	at	0.0007	4	a	0.0001
5	one	0.0005	5	none	0.0000

Step 6

B	Token	Prob	A	Token	Prob
1	chair	0.8830	1	chair	0.9996
2	white	0.0251	2	visible	0.0001
3	person	0.0153	3	actual	0.0001
4	fol	0.0147	4	empty	0.0000
5	broken	0.0050	5	Chair	0.0000

Figure 13: Case 5 from the POPE popular split (COCO subset).