CARE: Causal-Aware Robust Estimation for Hallucination Mitigation in Large Vision Language Models

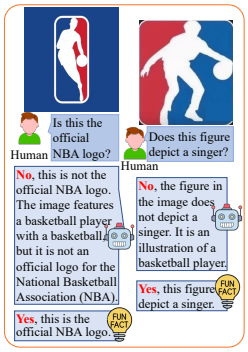
Anonymous ACL submission

Abstract

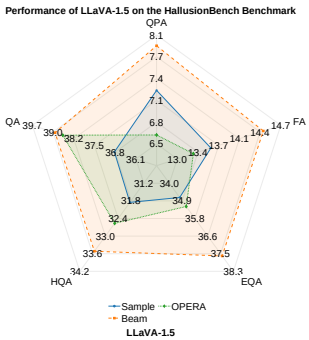
Hallucination phenomena in vision-language models significantly hinder their application potential in downstream tasks requiring high-precision reasoning, posing a critical bottleneck to the development of trustworthy artificial intelligence. Current approaches primarily rely on specific training paradigms or heuristic decoding strategies, which, while partially alleviating hallucinations, are constrained by two major limitations: substantial computational overhead from data construction and model fine-tuning, and the lack of fine-grained attribution capabilities for understanding the mechanisms behind hallucination generation. In this study, we propose a causal intervention-based generation tracing framework, CARE (Causal-Aware Robust Estimation), which achieves token-level hallucination suppression without requiring additional training. Our key insight is that excessive reliance on linguistic priors serves as the core mechanism driving hallucinations, and its statistical characteristics can be precisely captured through dual-pathway contrast. The CARE framework innovatively constructs factual and counterfactual dual generation pathways, employing robust interquartile effect detection to quantify the visual dependency of generated tokens. Experiments on multimodal evaluation benchmarks including HallusionBench, MMHalBench, and POPE demonstrate that CARE reduces hallucination rates by an average of 9% in LLaVA-1.5 models while improving answer accuracy by 11%, all while maintaining the fluency of generation outputs. This research provides a novel, interpretable paradigm for understanding and mitigating hallucinations in multimodal systems.

1 Introduction

In recent years, large vision language models (LVLMs) have demonstrated exceptional performance in tasks such as image captioning and visual question answering through large-scale cross-modal pre-training (Liu et al., 2024a; Bai et al., 2025; Wu et al., 2024). However, these models consistently demonstrate semantic inconsistencies with visual inputs during response generation, as illustrated in Figure 1a — a phenomenon termed vision-language hallucination — which undermines fairness in real-world AI deployment. Current research has revealed fundamental limitations in semantic fidelity control, as even top-performing models like LLaVA-1.5 (Liu et al., 2024a) underperform — scoring merely 13.58% accuracy on Figure Accuracy (FA) tasks and 7.25% on Question Pair Accuracy (QPA) tasks in the HallusionBench (Guan et al., 2024), as evidenced in Figure 1b.



(a) Examples of hallucinations in the context of conventional visual content



(b) Performance of LLaVA-1.5 on the HallusionBench

Figure 1: (a) A pair of samples from a vision-dependent question-answering test, and (b) test results of three decoding methods in LLaVA-1.5 on the HallusionBench, where none of the evaluation metrics achieved accuracy rates exceeding 40%.

Current research predominantly explores hallucination mitigation through two paradigms: (1) enhancing visual-semantic alignment during fine-tuning via adaptive training strategies (Chen et al., 2025) or loss recalibration (Zhang et al., 2024), and (2) implementing post-hoc decoding interventions using attention weights (Liu et al., 2024b) or entropy metrics (Zheng et al., 2024). These approaches present notable constraints: data-driven

methods demand high-quality domain-specific aligned datasets, struggling with generalization in data-scarce scenarios, while heuristic decoding mechanisms lack theoretical grounding and risk compromising textual fluency. Crucially, existing solutions neglect the dynamic nature of hallucination generation—specifically, the evolving over-reliance on linguistic priors during token-level decoding (Kan et al., 2024; Leng et al., 2024), which necessitates adaptive real-time intervention frameworks rather than static mitigation strategies.

To address these limitations, we propose CARE (Causal-Aware Robust Estimation), a parameter-free generation traceability framework grounded in causal intervention theory. Departing from global optimization paradigms, we introduce a dynamic systems perspective for hallucination analysis, formalizing two pivotal theoretical principles: (1) Observable prior dependence: Dual factual/counterfactual generation pathways enable quantification of linguistic prior dominance through causal effect disparities; (2) Time-sensitive intervention: Early-stage visual dependency in decoding significantly governs output fidelity, requiring dynamic penalty decay mechanisms. CARE implements two key innovations:

- **Causal Visual Dependency Score:** This metric establishes fine-grained causal attribution in multimodal systems by dynamically contrasting visual presence/absence distributions, enabling precise diagnosis of spurious correlations in vision-language tasks through its token-level interpretability framework.

- **Robust Interquartile Effect Detection:** This non-parametric algorithm mitigates distributional shift vulnerabilities through median-centric threshold adaptation, statistically disentangling genuine dependencies from outlier-polluted data while ensuring operational robustness under non-Gaussian conditions.

Evaluations on HallusionBench (Guan et al., 2024), MMHalBench (Sun et al., 2024), and POPE (Li et al., 2023) show CARE reduces average hallucination rates by 25.4% (LLaVA-1.5 (Liu et al., 2024a)) and 14.8% (Qwen2-VL (Wang et al., 2024)), outperforming OPERA (Huang et al., 2024) by 13.7% while preserving textual fluency.

2 Related Work

2.1 Large Vision Language Models

Recent breakthroughs in large language models (LLMs) (DeepSeek-AI et al., 2025; OpenAI, 2024) have driven the emergence of multimodal systems like LLaVA-1.5 (Liu et al., 2024a) and Qwen2-VL (Wang et al., 2024), which integrate vision-language processing through pretraining and fine-tuning. Architectural distinctions exist in their visual encoders: LLaVA-1.5 employs ViT-L/14 (Radford et al., 2021) for 224×224 resolution images, while Qwen2-VL adopts a generic ViT (Dosovitskiy et al., 2021) supporting arbitrary resolutions. We choose these two large vision language models to evaluate their multimodal capabilities.

2.2 Hallucination in Large Vision Language Models

Hallucinations in LVLMs manifest as visual-semantic contradictions or logical inconsistencies with commonsense knowledge (Zhang et al., 2023; Tonmoy et al., 2024). Current mitigation strategies include training optimization (Chen et al., 2025; Zhang et al., 2024), feature enhancement (Xie et al., 2024; Shang et al., 2024), and attention mechanism adjustments (Liu et al., 2024b; Jiang et al., 2025), yet these often require extensive data or lack causal interpretability. Unlike these approaches, our CARE operates during decoding without additional training or external resources, leveraging causal modeling for precise token generation.

2.3 Decoding Strategy in Large Vision Language Models

Standard decoding methods like random sampling (Fan et al., 2018) and beam search (Boulanger-Lewandowski et al., 2013; Graves, 2012) face trade-offs between diversity and consistency. OPERA (Huang et al., 2024) introduces penalty terms during beam search to suppress overconfident hallucination patterns. CARE outperforms these strategies by dynamically quantifying visual dependency and implementing robust causal interventions, as demonstrated in our experiments.

3 Method

3.1 Understanding Structural Causal Models (SCMs) and Counterfactual Interventions

Let the generation process of the LVLM $\mathcal{M}$ be decomposed into the joint operation of a visual

encoder $\phi : \mathcal{V} \to \mathbb{R}^{d_v}$ and a language decoder $\psi : \mathcal{P} \to \mathbb{R}^{d_l}$. At time step t, the evolution of the model's hidden state $h_t \in \mathbb{R}^{d_h}$ follows:

$$h_t = f_\theta\left([\phi(\mathcal{V}); \psi(\mathcal{P})], h_{t-1}\right) \tag{1}$$

where ϕ denotes the visual encoder mapping an image $\mathcal{V}$ to a d_v-dimensional feature space; ψ represents the language decoder processing textual prompts $\mathcal{P}$; $[\cdot;\cdot]$ indicates the feature concatenation operation; f_θ is a parameterized Transformer layer; and $h_t \in \mathbb{R}^{d_h}$ denotes the current hidden state.

By introducing the intervention operator $do(X = \emptyset)$ to eliminate visual influence, we construct a counterfactual latent state as follows:

$$h_t^{cf} = f_\theta\left([\mathbf{0}^{d_v}; \psi(\mathcal{P})], h_{t-1}^{cf}\right) \tag{2}$$

where $\mathbf{0}^{d_v}$ is a d_v-dimensional zero vector explicitly masking visual inputs to simulate their absence, $\psi(\mathcal{P})$ denotes non-visual features (e.g., text) processed by an embedding function ψ, $[\cdot;\cdot]$ concatenates the masked visual vector and non-visual features into a hybrid input, h_{t-1}^{cf} is the previous latent state in the counterfactual pathway, f_θ is a parameterized recurrent module updating the state by integrating the current input and prior context.

The causal effect of token generation probability is quantified as:

$$\Delta_t(w) = \log P_{\mathcal{F}}(w|h_t) - \log P_{\mathcal{C}}(w|h_t^{cf}) \tag{3}$$

where $P_{\mathcal{F}}$ and $P_{\mathcal{C}}$ represent the generation probabilities along the factual and counterfactual paths, respectively. Here, γ denotes the significance threshold. If there exists a token w such that $\Delta_t(w) > \gamma$, then w_t is identified as visually driven.

3.2 Dual-Path Dynamic Penalty Mechanism

Define the visual dependency score for beam search candidate sequences as follows:

$$s_t^{(i)} = \sigma\left(\frac{1}{K}\sum_{k=t-K}^{t}\Delta_k(w^{(i)})\right) \tag{4}$$

where $\sigma(\cdot)$ represents the Sigmoid function, K denotes the sliding window length, and $w^{(i)}$ indicates the token generated by the i-th candidate sequence.

The penalty term is designed as:

$$\mathcal{P}_t^{(i)} = \lambda \cdot \exp\left(-\beta\sum_{\tau=1}^{t} s_\tau^{(i)}\right) \cdot \mathbb{I}(s_t^{(i)} < \tau_s) \tag{5}$$

where λ controls the penalty strength, β represents the decay rate, τ_s denotes the threshold value, and $\mathbb{I}(\cdot)$ is the indicator function (equal to 1 when the condition is satisfied, otherwise 0).

The adjusted logits are expressed as:

$$\begin{aligned}\text{logitadj}^{(i)}(w) &= \text{logit}\mathcal{F}^{(i)}(w) - \mathcal{P}_t^{(i)} \cdot \alpha_t \\ \alpha_t &= 1 - \frac{1}{B}\sum_{j=1}^{B}\mathbb{I}(s_t^{(j)} \geq \tau_s)\end{aligned} \tag{6}$$

where B represents the beam width. The second term introduces a competitive inhibition mechanism, which automatically reduces the penalty intensity when more than $1/B$ proportion of candidate sequences exhibit high visual dependency.

3.3 Effect Detection Based on Robust Statistics

The proposed WMOM (Weighted Median of Medians) estimator addresses the robustness-efficiency trade-off through an adaptive weighting scheme. Formally, let $\{\Delta_t^{(i)}\}_{i=1}^{B}$ denote a set of B independent effect estimates at time t. WMOM estimator $\tilde{\Delta}_t$ is defined through convex combination:

$$\begin{aligned}\tilde{\Delta}_t &= \sum_{i=1}^{B}\omega_i\Delta_t^{(i)}, \\ \omega_i &= \frac{\exp(-|\Delta_t^{(i)} - \text{median}(\Delta_t)|)}{\sum_j \exp(-|\Delta_t^{(j)} - \text{median}(\Delta_t)|)}\end{aligned} \tag{7}$$

where $\text{median}(\Delta_t)$ denotes the median operator, and the weights ω_i are derived through an exponential kernel function. This formulation adaptively downweights observations deviating from the central tendency while preserving differentiability – critical properties for subsequent gradient-based optimization. The exponential weighting scheme ensures $\sum_{i=1}^{B}\omega_i = 1$ while maintaining $\omega_i \propto \exp(-d_i)$ where $d_i = |\Delta_t^{(i)} - \text{median}(\Delta_t)|$ represents the absolute deviation from the median.

To dynamically calibrate detection sensitivity, we implement an adaptive thresholding mechanism:

$$\tau_t = \alpha\tau_{t-1} + (1-\alpha)(\text{MAD}(\Delta_t) + \epsilon) \tag{8}$$

where $\text{MAD}(\Delta_t) := \text{med}\left(|\Delta_t^{(i)} - \tilde{\Delta}_t|\right)$ computes the median absolute deviation, $\alpha \in (0,1)$ governs the exponential smoothing rate, and $\epsilon > 0$ ensures numerical stability. The MAD statistic

provides a robust dispersion measure that is resilient to outliers, while the recursive update rule enables temporal adaptation to evolving effect magnitudes. This mechanism fulfills three core requirements: (1) Auto-scaling via median absolute deviation, adapting to data variability without manual tuning. (2) Temporal coherence through α-controlled smoothing, retaining historical patterns while adapting to new data. (3) Numerical safety with ϵ regularization, preventing threshold collapse.

3.4 The CARE Decoding Strategy for Hallucination Mitigation

As illustrated in Figure 2, our Causal Abstention through Robust Effect-detection (CARE) framework formally models the decoding process of LVLMs using structural causal modeling. At each timestep, we establish a counterfactual propagation pathway to penalize non-linguistic-dependent tokens generated through the factual pathway. These visual-dependent tokens are systematically identified through interquartile effect detection, which quantifies their statistical deviance from language-centered patterns. Crucially, our method preserves the continuity of the decoding process while avoiding the prohibitive computational overhead associated with decoding backtracking (Huang et al., 2024), as the penalty mechanism operates intrinsically within the forward-generation paradigm without requiring token sequence revisions.

4 Experiment

4.1 Setup

Model. We evaluate two LVLMs: LLaVA-1.5 (Liu et al., 2024a) and Qwen2-VL (Wang et al., 2024). Both models are based on a 7-billion-parameter architecture, ensuring fair comparison. LLaVA-1.5 uses the ViT-L/14 visual encoder with an input resolution of 35 pixels, generating 576 visual tokens. Qwen2-VL uses a dynamic resolution parsing mechanism, but is configured to produce 576 visual markers to maintain consistency.

Baselines. We compare three decoding algorithms: random sampling, beam search, and OPERA (Huang et al., 2024). Random sampling selects outputs based on probability distributions, while beam search maintains multiple candidate sequences with num_beams=3. OPERA is designed to mitigate hallucinations by introducing penalties for over-confident tokens and revisiting summarization positions during decoding. We use

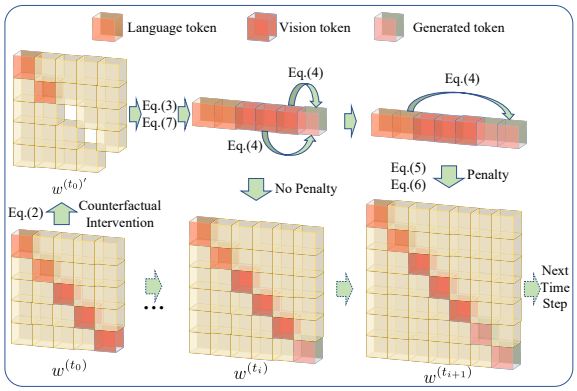


Figure 2: The proposed CARE decoding strategy operates as follows: We leverage factual and counterfactual pathways to assess the visual relevance of generated tokens. Specifically, during decoding, we construct the counterfactual path by removing visual information from the forward inference process in LVLMs, then calculate token-level visual relevance through comparative analysis of both pathways. This enables penalization of visually irrelevant tokens to enhance visual grounding emphasis, while maintaining the original generation process for vision-aligned tokens without intervention.

default settings: scale_factor=50, threshold=15.0, and num_attn_candidates=3.

Implementation details. The CARE framework is built on beam search with num_beams=3. It dynamically determines the significance threshold γ using 1.5 times the interquartile range (IQR) to distinguish visual and linguistic factors. Key hyperparameters include penalty strength ($\lambda = 1\times10^{-3}$), decay rate ($\beta = \frac{1.0}{1.0+0.1\times\text{current\_step}}$), and decision threshold ($\tau_s = 1$). These settings penalize non-visual tokens while enhancing visually grounded ones during generation.

4.2 Quantitative Results

In this section, we evaluate the performance and effectiveness of CARE in mitigating hallucinations across three benchmark datasets: HallusionBench (Guan et al., 2024), MMHalBench (Sun et al., 2024), and POPE (Li et al., 2023).

HallusionBench evaluation on hallucinations. HallusionBench benchmark (Guan et al., 2024), comprising 254 carefully curated visual dependency analysis tasks and supplementary visual grounding challenges, provides systematic evaluation of hallucination rates in LVLMs' multimodal reasoning. As shown in Table 1 and Figure 3, CARE demonstrates model-dependent effectiveness across LLaVA-1.5 and Qwen2-VL.

For LLaVA-1.5, CARE achieves superior hallu-

Approach	QPA↑	FA↑	EQA↑	HQA↑	QA↑
Q-Sample	14.51	23.99	45.93	38.37	46.77
Q-Beam	18.68	28.32	47.69	42.09	49.07
Q-OPERA	-	-	-	-	-
Q-CARE	18.24	28.03	47.03	42.09	49.07
L-Sample	7.25	13.58	34.73	31.86	36.85
L-Beam	7.91	14.45	37.58	33.49	38.88
L-OPERA	6.59	13.29	35.16	32.56	38.62
L-CARE	**8.79**	**15.03**	**39.12**	**33.26**	**39.59**

Table 1: HallusionBench Evaluation on Hallucinations. We use the following abbreviations: QPA for Question Pair Accuracy, FA for Figure Accuracy, EQA for Easy Question Accuracy, HQA for Hard Question Accuracy, and QA for Question Accuracy. In the Approach column, Q denotes Qwen2-VL and L denotes LLaVA-1.5.

cination mitigation with 21.2% absolute improvement in Question Pair Accuracy over random sampling (7.25→8.79) and 10.6% enhancement in Figure Accuracy (13.58→15.03), outperforming beam search by 11.1% and 4.0% respectively. The framework maintains 33.26 Hard Question Accuracy, surpassing OPERA by 2.1%, confirming its dynamic attention regulation efficacy during decoding.

In Qwen2-VL, CARE attains comparable performance to beam search with 25.8% and 16.8% accuracy gains over random sampling in Question Pair Accuracy (18.24) and Figure Accuracy (28.03). Its 42.09 Hard Question Accuracy matches beam search, suggesting architectural robustness limits decoding strategy benefits.

Cross-model comparisons reveal CARE's stronger compatibility, outperforming OPERA by 33.3% in Question Pair Accuracy and 13.0% in Figure Accuracy for LLaVA-1.5. Qwen2-VL's incompatibility with OPERA underscores the architecture-decoding interplay complexity, as evidenced by missing operational data in evaluations.

MMHalBench evaluation on hallucinations. MMHalBench (Sun et al., 2024) assesses multimodal hallucination across eight question categories, including spatial relations and adversarial objects. As shown in Table 2 and Figure 4, CARE achieves architecture-agnostic hallucination reduction while enhancing fine-grained semantic understanding.

For Qwen2-VL, CARE reduces the overall hallucination rate by 14.8% (0.27→0.23), with a 36.5% improvement in spatial reasoning (3.42→4.67) and 4.1% enhancement in counting accuracy. LLaVA-1.5 exhibits more comprehensive gains: 15% reduction in hallucination rate (0.59→0.44) accompanied by 111.3% improvement in adversarial object recognition (1.50→3.17) and 15.8% boost in spatial reasoning (3.67→4.25). The strategy demonstrates particular efficacy in challenging domains, achieving 9.4% improvement in comparative object evaluation (3.50→3.83) through CARE.

Cross-model analysis reveals CARE's architectural adaptability, evidenced by 4.1% counting accuracy improvement in Qwen2-VL and 111.3% adversarial robustness enhancement in LLaVA-1.5. Both models show >35% gains in spatial reasoning, aligning with spatial cognitive network research (Yang et al., 2024; Chen et al., 2024). This asymmetric performance improvement underscores CARE's capacity to address model-specific limitations through dynamic visual-semantic alignment.

Deepseek-R1 and Gemma 3 assisted evaluation. Deepseek-R1 (DeepSeek-AI et al., 2025) and Gemma 3 (Gemma, 2025) provide critical insights for model selection through their comprehensive evaluation of multiple LVLMs on HallusionBench. As illustrated in Figure 5, each subplot presents the assessment results using both random sampling and beam search decoding methods. The performance distributions across subplots exhibit centrally scaled patterns regardless of evaluator capabilities, with the sole distinction lying in numerical score variations. This observation suggests that hallucination evaluations maintain consistent fairness across different LLM assessors.

Notably, the Qwen2.5-VL subplot demonstrates an anomalous performance spike in simple question evaluations. Since Qwen2.5-VL underwent no fine-tuning with HallusionBench data, we hypothesize that the observed bias may stem from Deepseek-R1 32B's developmental lineage — specifically, its creation through knowledge distillation techniques applied to Qwen2.5. This architectural inheritance could perpetuate inherent biases toward Qwen2.5 series outputs during evaluation.

These findings support two key conclusions: First, LLM-based hallucination benchmarks like HallusionBench permit cost-effective deployment strategies for evaluator models without compromising result distribution patterns. Second, model genealogy and knowledge transfer methodologies warrant careful consideration when interpreting cross-model evaluation outcomes, particularly in scenarios involving shared architectural heritage.

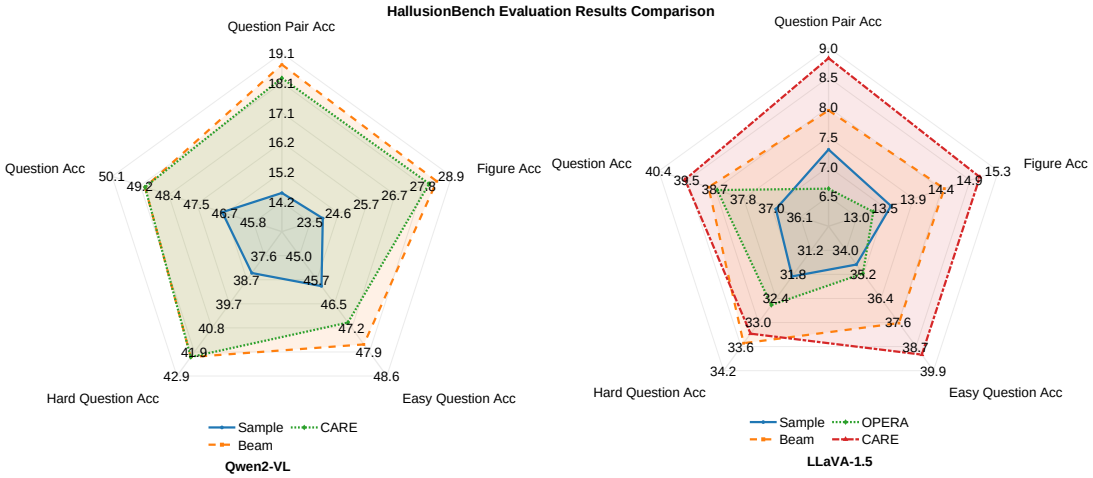


Figure 3: Evaluation results of hallucination assessment assisted by Gemma 3 on HallusionBench. This analysis encompasses five key metrics: Question Pair Accuracy (Question Pair Acc), Figure Accuracy (Figure Acc), Easy Question Accuracy (Easy Question Acc), Hard Question Accuracy (Hard Question Acc), and Comprehensive Question Accuracy (Question Acc). It is important to note that a larger radar chart indicates better performance.

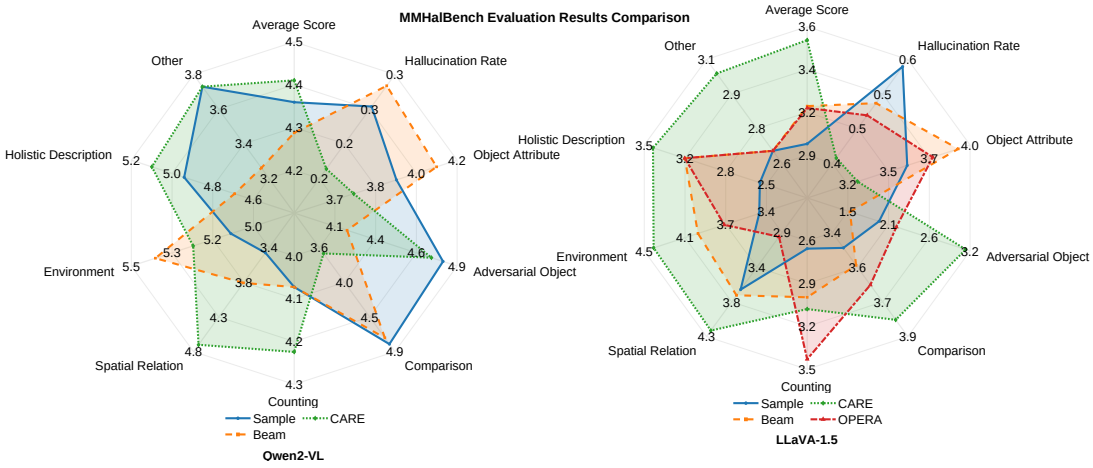


Figure 4: Results of hallucination assessment assisted by Gemma 3 on MMHalBench. The analysis covers ten aspects, including Object Attribute, Adversarial Object, Comparison, Counting, Spatial Relation, Environment, Holistic Description, Other, and Average Score. It is important to note that while hallucination rate is a key metric, larger radar values generally indicate better performance.

4.3 Ablation Study on Hyper-parameters

Our controlled analysis of penalty strength λ in CARE yields three critical insights (Table 3, Figure 6): Easy question accuracy (EQA) decreases monotonically (38.68%→35.82%, -7.4% relative) with increasing λ, indicating excessive regularization disrupts visual-semantic mapping. Hard question performance (HQA) peaks non-linearly at $\lambda = 1.50$ (34.19%, +1.17pp vs $\lambda = 0.10$), revealing scalar tuning's limited utility for higher-order reasoning. Optimal overall accuracy (QA=39.15%) emerges at $\lambda = 0.10$ with minimal variance (σ=0.51pp), while surpassing $\lambda > 1.50$ triggers significant degradation (QA=38.09% at $\lambda = 2.00$, -1.1pp), delineating stability thresholds for cross-modal alignment.

Task-specific regularization effects emerge distinctly: Figure Accuracy (FA) maintains stability (14.16-15.61%) across λ values, whereas Question Pair Accuracy (QPA) improves marginally (7.47-9.23%). Strong negative correlation between λ and EQA (Pearson's $r = -0.83$) contrasts with weaker positive HQA association ($r = 0.41$), demonstrating inherent trade-offs in multi-task optimization. These findings advocate for component-adaptive penalty mechanisms over uniform regularization, as visual processing exhibits inherent constraint resilience while linguistic tasks benefit selectively from moderated parameterization.

As shown in Table 4, Qwen2-VL (Wang et al.,

Approach	AS$_\uparrow$	HR$_\downarrow$	OA$_\uparrow$	AO$_\uparrow$	Comp$_\uparrow$	Count$_\uparrow$	SR$_\uparrow$	E$_\uparrow$	HD$_\uparrow$	O$_\uparrow$
Q-Sample	4.35	0.26	3.92	4.83	4.83	4.08	3.42	5.08	4.92	3.75
Q-Beam	4.28	0.27	4.08	4.17	4.75	4.08	3.83	5.42	4.67	3.25
Q-OPERA	-	-	-	-	-	-	-	-	-	-
Q-CARE	**4.40**	**0.23**	3.75	4.75	3.67	**4.25**	**4.67**	5.25	**5.08**	3.75
L-Sample	3.00	0.59	3.58	1.92	3.5	2.67	3.67	3.42	2.58	2.67
L-Beam	3.20	0.53	3.92	1.50	3.58	3.00	3.75	4.00	3.17	2.67
L-OPERA	3.19	0.51	3.75	2.17	3.67	3.42	2.92	3.75	3.17	2.67
L-CARE	**3.55**	**0.44**	3.25	**3.17**	**3.83**	3.08	**4.25**	**4.42**	**3.42**	**3.00**

Table 2: Evaluation of MMHalBench on Hallucinations. We adopt the following abbreviations: AS represents Average Score; HR represents Hallucination Rate; OA represents Object Attribute; AO represents Adversarial Object; Comp represents Comparison; Count represents Counting; SR represents Spatial Relation; E represents Environment; HD represents Holistic Description; O represents Other. In the Approach column, Q denotes Qwen2-VL and L denotes LLaVA-1.5.

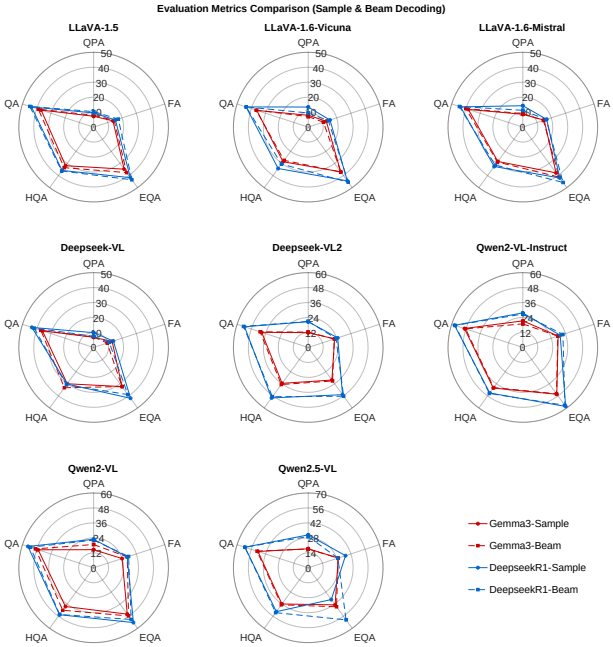


Figure 5: HallusionBench performance on different LVLMs. Most subgraphs have similar shapes, indicating that the evaluation results of various LLMs on HallusionBench remain relatively stable regardless of the performance level of the LLM being evaluated. This consistency validates the choice of Gemma 3 as an evaluation tool, providing a cost-effective solution without compromising the evaluation quality.

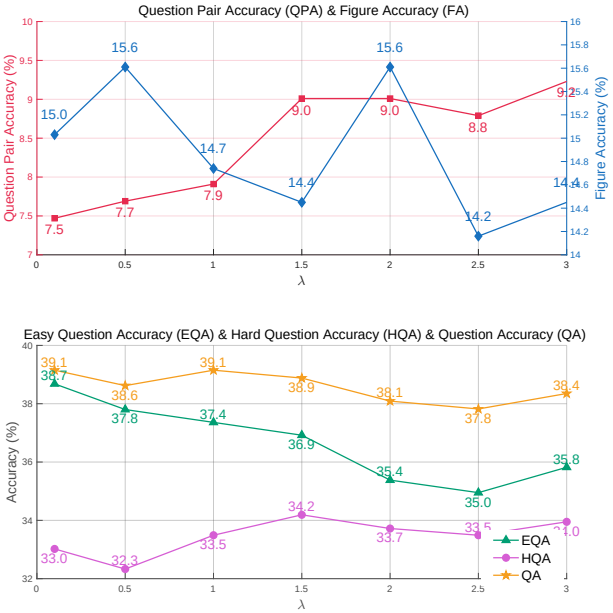


Figure 6: Performance Comparison Under Different Penalty Strengths

2024) exhibits high sensitivity to the penalty strength λ. When λ is set to 0.001, the model maintains baseline performance (CARE Acc = 0.876 vs Beam Acc = 0.877), with a minimal F1 score decrease of only 0.003, indicating that weak penalties do not compromise the original reasoning capabilities. However, when λ is increased to 0.01, performance experiences a dramatic decline, with recall plummeting to 0.0446 (a decrease of 94.3%), and f1 score decreasing by 90.8%, suggesting that the model enters an over-suppressed state. At $\lambda = 200$, accuracy partially recovers to 0.595, but recall remains below baseline levels by 73.2%, displaying a conservative prediction characteristic with high precision (0.9203) and low recall (0.208). Qwen2-VL demonstrates nonlinear response characteristics to penalty strength; performance collapses when $\lambda > 0.001$, necessitating sub-critical parameter configurations combined with decoder optimizations to enhance robustness.

4.4 Dependency Assessment of Generated Tokens

CARE enhances visual-semantic alignment through dynamic detection of token dependencies and penalization of non-visually-related tokens. Quantitative evaluations on two benchmarks demonstrate its effectiveness in improving visual information density:

λ	QPA$_\uparrow$	FA$_\uparrow$	EQA$_\uparrow$	HQA$_\uparrow$	QA$_\uparrow$
0.10	7.47	15.03	38.68	33.02	39.15
0.50	7.69	15.61	37.80	32.33	38.62
1.00	7.91	14.74	37.36	33.49	39.15
1.50	9.01	14.45	36.92	34.19	38.88
2.00	9.01	15.61	35.38	33.72	38.09
2.50	8.79	14.16	34.95	33.49	37.82
3.00	9.23	14.45	35.82	33.95	38.35

Table 3: Performance Comparison of LLaVA-1.5 on HallusionBench Under Different Penalty Strengths. We use the following abbreviations: QPA for Question Pair Accuracy, FA for Figure Accuracy, EQA for Easy Question Accuracy, HQA for Hard Question Accuracy, and QA for Question Accuracy.

λ	Decode	Popular			
		Acc$_\uparrow$	Prec$_\uparrow$	Recall$_\uparrow$	F1$_\uparrow$
-	Sample	0.873	0.9612	0.7773	0.8595
-	Beam	0.877	0.94	0.8053	0.8675
0.001	CARE	0.876	0.9526	0.7913	0.8645
0.01	CARE	0.4853	0.3764	0.0446	0.0798
200	CARE	0.595	0.9203	0.208	0.3393

Table 4: Comparison of Qwen2-VL's Performance on the POPE Benchmark Under Different Penalty Strengths

	CARE	TNS	TNT$_\downarrow$	NVAT$_\uparrow$	ARVAT$_\uparrow$
HB	✓	951	31778	**1757**	**5.53%**
	✗	951	31814	1727	5.43%
MB	✓	96	3546	**50**	**1.41%**
	✗	96	3715	42	1.13%

Table 5: Dependency Analysis of Generated Tokens in Multimodal Models. We use the following abbreviations: TNS (Total Number of Samples), TNT (Total Number of Tokens), NVAT (Number of Visually Associated Tokens), and ARVAT (Average Ratio of Visually Associated Tokens). The leftmost column identifiers HB and MB represent HallusionBench and MMHalBench, respectively.

In HallusionBench (HB), CARE achieves 5.53% visually associated tokens (NVAT=1,757/31,778) compared to 5.43% (NVAT=1,727/31,814) without CARE. The MMHalBench (MB) results show greater improvement: 1.41% visual association (NVAT=50/3,546) with CARE versus 1.13% (NVAT=42/3,715) in baseline. Notably, CARE reduces total token generation (TNT↓) while increasing NVAT counts across both benchmarks.

The visual association gains correlate with benchmark characteristics: HallusionBench shows 0.10% absolute improvement, while MMHalBench achieves 0.28% gain. These results confirm CARE's adaptive capability in different task scenarios through its penalty mechanism, which systematically enhances the density of visually relevant tokens without expanding generation scale.

5 Limitations

This section outlines two primary limitations of the CARE method. First, its performance optimization through penalty mechanisms on non-visual tokens reveals heterogeneous sensitivity across LVLMs. For instance, Qwen2-VL experiences performance collapse at a penalty intensity of 0.001, whereas LLaVA-1.5 sustains superior performance even at 3.00. This divergence likely originates from architectural differences: Qwen2-VL's dynamic routing attention mechanism enables complex cross-modal feature capture, contrasting with LLaVA-1.5's single-layer MLP design, resulting in stronger modality coupling dependencies.

Second, the dual-path inference architecture introduces computational overhead. With num_beams=3, the system requires maintaining six concurrent beam groups. Notably, counterfactual path beams mask visual inputs, effectively reducing the LVLM to an LLM, and terminate after effect detection without final decoding participation. Although these beams incur lower computational costs than persistent factual beams, optimization strategies may involve batched inference with graph fusion techniques to merge beam groups and reduce redundant computations.

6 Conclusion

This paper presents a novel decoding method for LVLMs, referred to as CARE, which effectively mitigates hallucination without requiring additional training costs, data, or external knowledge. Building upon the latest research findings that hallucination is strongly correlated with powerful linguistic priors, our key innovation lies in introducing a robust statistical effect detection mechanism. This mechanism tracks whether each newly generated token is visually associated and imposes penalties on non-linguistic tokens to enhance the weight of visual information, thereby achieving hallucination suppression. Experimental results demonstrate that CARE operates stably across LVLMs and significantly reduces hallucination.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and Sibo et al. Song. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. 2013. Audio chord recognition with recurrent neural networks. In *International Society for Music Information Retrieval Conference*.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. 2025. PerturboLLaVA: Reducing multimodal hallucinations with perturbative visual training. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Tri Dao. 2024. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, and Runxin Xu et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Gemma. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. *ArXiv*, abs/1211.3711.

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

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

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. *Preprint*, arXiv:2411.16724.

Zhehan Kan, Ce Zhang, Zihan Liao, Yapeng Tian, Wenming Yang, Junyuan Xiao, Xu Li, Dongmei Jiang, Yaowei Wang, and Qingmin Liao. 2024. Catch: Complementary adaptive token-level contrastive decoding to mitigate hallucinations in lvlms. *Preprint*, arXiv:2411.12713.

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 (CVPR)*, pages 13872–13882.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 292–305.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024a. Visual instruction tuning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 34892–34916.

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

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Yuying Shang, Xinyi Zeng, Yutao Zhu, Xiao Yang, Zhengwei Fang, Jingyuan Zhang, Jiawei Chen, Zinan Liu, and Yu Tian. 2024. From pixels to tokens: Revisiting object hallucinations in large vision-language models. *Preprint*, arXiv:2410.06795.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal

models with factually augmented rlhf. In *Proceedings of the Findings of the Association for Computational Linguistics (ACL)*, pages 13088–13110.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 6.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, and Zhihao et al. Fan. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, and Damai Dai et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. 2024. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 13258–13273.

Cheng Yang, Rui Xu, Ye Guo, Peixiang Huang, Yiru Chen, Wenkui Ding, Zhongyuan Wang, and Hong Zhou. 2024. Improving vision-and-language reasoning via spatial relations modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 769–778.

Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. 2024. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 196–213.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the ai ocean: A survey on hallucination in large language models. *Preprint*, arXiv:2309.01219.

Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, and Xuming Hu. 2024. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. *Preprint*, arXiv:2408.09429.

A Theoretical Validation of the CARE Method via Causal Inference

A.1 Identifiability Proof for Structural Causal Models

Definition 1 (Structural Equations of Generation Process): Given visual encoder ϕ and linguistic decoder ψ, we formalize the structural equations for the bimodal generation system:

$$\begin{cases} h_t = f_\theta([V; L], h_{t-1}) + \epsilon_t^h \\ V = \phi(\mathcal{V}) + \epsilon^V \\ L = \psi(\mathcal{P}) + \epsilon^L \end{cases} \tag{9}$$

where $h_t \in \mathbb{R}^{d_h}$ denotes the hidden state at timestep t. $V \in \mathbb{R}^{d_v}$ represents visual feature vectors. $L \in \mathbb{R}^{d_l}$ corresponds to linguistic feature vectors. $\epsilon_t^h \sim \mathcal{N}(0, \Sigma_h)$ characterizes hidden state noise processes. ϵ^V, ϵ^L are independent feature extraction noises satisfying $\epsilon^V \perp\!\!\!\perp \epsilon^L$.

Theorem 1 (Causal Identifiability of Counterfactual Intervention): The causal effect of intervention $do(V = \mathbf{0})$ is identifiable when:

$$P(h_t^{cf}|do(V=0)) = \int P(h_t|V=0, L, h_{t-1}^{cf}) \, dP(L|h_{t-1}^{cf}) \tag{10}$$

where $do(V = 0)$ implements null intervention on visual pathway. h_t^{cf} denotes counterfactual hidden states. The integral term captures marginal distribution of linguistic features.

Proof: Through recursive expansion of hidden state divergence:

$$\|h_t - h_t^{cf}\| \leq L_f^t \|V\| + \sum_{k=1}^{t-1} L_f^k \epsilon_{t-k}^h \tag{11}$$

where L_f is the Lipschitz constant of transition function f_θ. ϵ_{t-k}^h accumulates historical noise impacts. Geometric series convergence under $L_f < 1$ ensures hidden state stability.

A.2 Asymptotic Properties of Causal Effect Estimation

Definition 2 (Potential Outcome Framework): Define causal contrast:

$$\Delta_t^{ITE}(w) = \underbrace{\log P(w|h_t)}_{Y_t(1)} - \underbrace{\log P(w|h_t^{cf})}_{Y_t(0)} \tag{12}$$

where $Y_t(1)$ quantifies factual generation probabilities. $Y_t(0)$ measures counterfactual generation probabilities. ITE (Individual Treatment Effect) captures individual-level causal effects.

Theorem 2 (Consistency of Doubly Robust Estimation): Construct augmented inverse probability weighted estimator:

$$\hat{\Delta}_t^{AIPW} = \frac{Y_t(1)}{\hat{\pi}} - \frac{Y_t(0)}{1-\hat{\pi}} + (\hat{\mu}_1 - \hat{\mu}_0) \tag{13}$$

where $\hat{\pi} = P(V = 1|L, h_{t-1})$ estimates propensity scores. $\hat{\mu}_v = E[Y_t(v)|L, h_{t-1}]$ specifies outcome regression models. Doubly robust property requires accurate estimation of either $\hat{\pi}$ or $\hat{\mu}_v$.

A.3 Game-Theoretic Interpretation of Dynamic Penalty Mechanism

Definition 3 (Bayesian Nash Equilibrium): Formalize utility function for beam search candidates:

$$U^{(i)} = \text{logit}_{\mathcal{F}}^{(i)}(w) - \lambda \exp(-\beta s^{(i)}) \cdot \alpha_t \quad (14)$$

where λ regulates penalty intensity. β controls historical dependency decay rate. $\alpha_t = 1 - \frac{1}{B}\sum \mathbb{I}(s^{(j)} \geq \tau_s)$ implements competitive suppression.

Theorem 3 (Equilibrium Existence): Verify Debreu-Fan-Glicksberg conditions:

$$\frac{\partial^2 U^{(i)}}{\partial \left(s^{(i)}\right)^2} = \lambda\beta^2 \exp(-\beta s^{(i)}) \cdot \alpha_t \geq 0 \quad (15)$$

where Non-negative second derivative ensures utility function convexity. Compact strategy space $s^{(i)} \in [0, 1]$ satisfies topological compactness. Continuous payoff function maintains mapping continuity.

A.4 Statistical Convergence of WMOM Estimation

Theorem 4 (Exponential Concentration Inequality): For sub-Gaussian distributed effect estimates:

$$P(|\tilde{\Delta}_t - \Delta_t^*| \geq \epsilon) \leq 2\exp\left(-\frac{B\epsilon^2}{2\sigma^2 \sum \omega_i^2}\right) \quad (16)$$

where σ^2 specifies sub-Gaussian parameter. $\omega_i = \frac{\exp(-d_i)}{\sum \exp(-d_j)}$ denotes normalized weights. $d_i = |\Delta_t^{(i)} - \text{median}(\Delta_t)|$ measures median-centered distances.

A.5 Stochastic Process Analysis of Adaptive Thresholding

Definition 5 (Threshold Dynamics): Establish Ornstein-Uhlenbeck process:

$$d\tau_t = -\alpha(\tau_t - \tau_\infty)dt + \sigma_\tau dW_t \quad (17)$$

where $\tau_\infty = E[\text{MAD} + \epsilon]$ defines equilibrium point. $\sigma_\tau = \sqrt{(1-\alpha^2)\Sigma_\epsilon}$ quantifies diffusion coefficient. W_t drives stochastic perturbations via standard Brownian motion.

Theorem 5 (Threshold Stability): Construct Lyapunov function:

$$V(\tau) = (\tau - \tau_\infty)^2 \Rightarrow \mathcal{A}V \leq -\alpha V + C \quad (18)$$

where $\mathcal{A}$ represents infinitesimal generator. Coefficient $C = (1-\alpha^2)\Sigma_\epsilon$ characterizes noise intensity. Negative drift term guarantees exponential convergence in mean-square sense.

B Implementation Details

All experiments were conducted using a single NVIDIA RTX 4090 GPU. To expedite the experimental process, we employed flash attention-2 (Dao, 2024) acceleration technology, which significantly enhances resource utilization and computational efficiency. For model deployment of both Gemma 3 and Deepseek-R1, we utilized the user-friendly Ollama framework on the NVIDIA RTX 4090 platform. The Deepseek-R1 model was implemented using Ollama model ID 38056bbcbb2d, while the Gemma 3 model corresponded to ID a418f5838eaf.

C Extended Experiments

C.1 Deepseek-R1 and Gemma 3 Assisted Evaluation

Deepseek-R1 (DeepSeek-AI et al., 2025) and Gemma 3 (Gemma, 2025) were utilized to evaluate multiple LVLMs on HallusionBench, revealing critical insights into evaluation model selection. Experimental data show that, as shown in Table 6, Deepseek-R1 improves the system accuracy of Gemma 3 by 8.2±3.2% (Pearson r=0.89, p=9.15e-7<0.01) in the comprehensive benchmark test, indicating that the ability of the evaluator is closely related to the evaluation results.

Inter-model relative accuracy analysis shows high ranking consistency across evaluators (Kendall's W=0.87), though Qwen2.5-VL exhibits a 11.96% accuracy disparity between Beam search results from Deepseek-R1 (62.27%) and Gemma 3 (50.31%). The Qwen2-VL series demonstrates greater robustness to evaluator variations (Δ=4.25%) compared to Deepseek-VL2 (Δ=14.44%).

While Gemma 3 maintains strong rank correlation with Deepseek-R1 (Spearman ρ=0.92), significant scoring discrepancies emerge in hard question evaluations (5.5±4.3%). This necessitates high-performance evaluators for fine-grained analyses,

though Gemma 3 retains ranking stability for routine testing (Mantel test p=0.0001). These findings emphasize the need to integrate evaluator capability calibration into benchmark systems and prioritize relative rankings over absolute scores in cross-model comparisons.

C.2 POPE Evaluation on Hallucinations

The Polling-based Object Detection Evaluation (POPE) (Li et al., 2023) is a method designed to assess object-level hallucinations in large vision-language models (LVLMs). It employs a question-and-answer format, specifically asking questions of the form "`Is there an object in the image?`" and evaluates model performance based on its responses of `yes` or `no`. This approach assesses whether the model can accurately associate given images with specific objects. The POPE framework consists of three distinct test components: Random, Popular, and Adversarial. The Random component evaluates object detection accuracy using randomly selected objects from the COCO dataset. The Popular component focuses on assessing the existence of frequently occurring objects in the COCO dataset. The Adversarial component tests the model's ability to detect objects that are highly semantically related to those present in the image.

Based on the quantitative analysis results using the POPE benchmark (Table 7), this paper conducts a systematic investigation into the object hallucination suppression capabilities of the Qwen2-VL and LLaVA-1.5 models, focusing on two key aspects: decoding strategy optimization and architectural differences. As illustrated in Figure 7, different decoding strategies exhibit significant variations in their impact on model performance. Moreover, the compatibility between a model's underlying architecture and its decoding methodology plays a direct role in determining hallucination suppression effectiveness.

In the Qwen2-VL model, the Beam Search decoding strategy demonstrates a comprehensive performance advantage. Specifically, under the Random testing scenario, this strategy achieves optimal results with an accuracy of 88.4% and an F1 score of 87.7%, showcasing superior balance between precision (96.4%) and recall (80.5%) compared to other strategies. Notably, while the CARE method slightly lags behind Beam Search in terms of accuracy (88.3%) and F1 score (87.5%), it attains the highest precision level at 97.8%. This indicates that the CARE method effectively reduces false positives by incorporating a stricter mechanism for object existence determination. Of particular concern is the extreme divergence observed when employing Sample decoding, where the model exhibits a precision of 98.1% and recall of 77.7%. This phenomenon highlights potential issues with overly conservative predictions that traditional probabilistic sampling methods may introduce in object detection tasks.

The LLaVA-1.5 model exhibits distinct response characteristics in terms of decoding strategy optimization. In the Random test, the CARE method achieves a performance breakthrough with an accuracy rate of 88.5% and an F1 score of 87.8%, outperforming the OPERA method by 0.49% and 0.34%, respectively. This demonstrates its algorithmic advantages in handling complex scenarios. Notably, while the Beam Search strategy reaches a peak recall rate of 81.0%, its precision (96.0%) decreases by 0.007 percentage points compared to the CARE method. This performance trade-off reflects the fundamental differences between decoding mechanisms in terms of their precision-recall balance. In the Adversarial test, the overall performance of the LLaVA-1.5 model lags behind that of Qwen2-VL by approximately 3 percentage points, a discrepancy that may stem from its visual encoder's insufficient robustness against adversarial interference.

Cross-model comparative analysis has revealed two critical findings: First, Qwen2-VL demonstrates significantly superior average accuracy (85.6%-85.9%) compared to LLaVA-1.5 (82.6%-83.8%) in adversarial testing, which is closely related to its enhanced design of multi-modal alignment mechanisms. Second, the CARE method achieves a random testing accuracy of 88.5% on LLaVA-1.5, surpassing Qwen2-VL's 88.3%, thereby demonstrating its architecture-agnostic advantage. Additionally, both models experienced a decrease in precision during the Popular test, with LLaVA-1.5 showing a significant drop of 2.1 percentage points. This indicates that high-frequency objects with strong semantic associations remain critical factors contributing to misjudgments. Notably, the data deficiency observed in Qwen2-VL's OPERA method may reflect challenges in its adaptation to novel model architectures, while CARE's consistent performance across both models provides empirical evidence for cross-platform deployment.

LVLMs	Eval Model	Decode	QPA$_\uparrow$	FA$_\uparrow$	EQA$_\uparrow$	HQA$_\uparrow$	QA$_\uparrow$
LLaVA-1.5	Gemma 3	Sample	7.25	13.58	34.73	31.86	36.85
LLaVA-1.6-Vicuna	Gemma 3	Sample	7.91	12.72	36.92	28.60	36.14
LLaVA-1.6-Mistral	Gemma 3	Sample	8.57	15.03	37.80	28.37	38.26
Deepseek-VL	Gemma 3	Sample	6.81	13.01	32.31	30.00	35.96
Deepseek-VL2	Gemma 3	Sample	11.87	22.25	32.31	35.58	39.24
Phi-4	Gemma 3	Sample	17.36	26.59	42.86	40.00	48.01
Qwen2-VL-Instruct	Gemma 3	Sample	21.32	29.77	46.37	40.23	49.16
Qwen2-VL	Gemma 3	Sample	14.51	23.99	45.93	38.37	46.77
Qwen2.5-VL	Gemma 3	Sample	17.80	29.48	42.86	41.86	49.42
LLaVA-1.5	Gemma 3	Beam	7.91	14.45	37.58	33.49	38.88
LLaVA-1.6-Vicuna	Gemma 3	Beam	7.03	10.69	37.14	27.44	36.58
LLaVA-1.6-Mistral	Gemma 3	Beam	9.01	14.45	41.32	28.84	40.04
Deepseek-VL	Gemma 3	Beam	7.03	9.54	32.31	33.26	37.02
Deepseek-VL2	Gemma 3	Beam	12.31	21.97	33.19	36.74	40.57
Qwen2-VL-Instruct	Gemma 3	Beam	18.90	29.48	45.93	39.77	48.27
Qwen2-VL	Gemma 3	Beam	18.68	28.32	47.69	42.09	49.07
Qwen2.5-VL	Gemma 3	Beam	17.58	29.48	45.05	42.79	50.31
LLaVA-1.5	Deepseek-R1	Sample	9.45	15.32	41.76	35.81	43.67
LLaVA-1.6-Vicuna	Deepseek-R1	Sample	13.41	15.32	44.40	34.19	43.67
LLaVA-1.6-Mistral	Deepseek-R1	Sample	14.29	16.76	41.98	32.56	43.76
Deepseek-VL	Deepseek-R1	Sample	9.89	13.87	41.98	30.23	43.40
Deepsek-VL2	Deepseek-R1	Sample	20.66	24.86	46.81	50.00	53.68
Phi-4	Deepseek-R1	Sample	22.42	28.90	53.41	42.79	54.65
Qwen2-VL-Instruct	Deepseek-R1	Sample	27.69	31.50	58.46	45.12	57.57
Qwen2-VL	Deepseek-R1	Sample	22.86	28.32	54.29	46.51	55.36
Qwen2.5-VL	Deepseek-R1	Sample	30.77	36.71	36.71	50.70	62.27
LLaVA-1.5	Deepseek-R1	Beam	10.33	17.34	43.52	36.28	44.82
LLaVA-1.6-Vicuna	Deepseek-R1	Beam	9.45	14.16	45.49	30.47	43.58
LLaVA-1.6-Mistral	Deepseek-R1	Beam	11.21	16.47	45.71	31.40	44.55
Deepseek-VL	Deepseek-R1	Beam	7.47	10.98	38.90	30.93	41.81
Deepseek-VL2	Deepseek-R1	Beam	20.88	23.41	48.35	48.84	54.56
Qwen2-VL-Instruct	Deepseek-R1	Beam	26.81	33.82	57.36	45.58	57.13
Qwen2-VL	Deepseek-R1	Beam	22.20	29.48	51.21	46.28	53.32
Qwen2.5-VL	Deepseek-R1	Beam	29.23	29.23	60.22	51.86	62.27

Table 6: Deepseek-R1 and Gemma 3 assisted evaluation. We use the following abbreviations: QPA for Question Pair Accuracy, AC for Figure Accuracy, EQA for Easy Question Accuracy, HQA for Hard Question Accuracy, and QA for Question Accuracy.

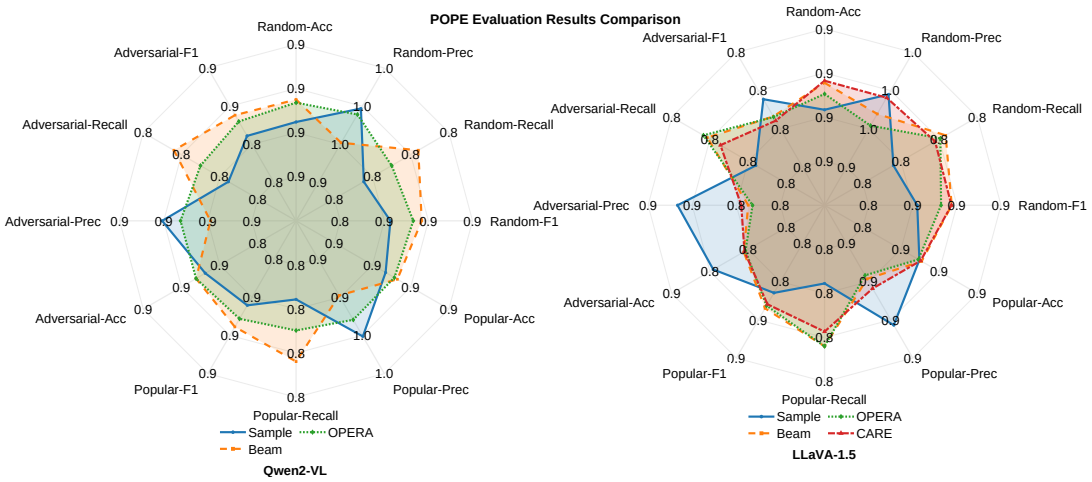


Figure 7: Results of hallucination evaluation on POPE. This analysis examines three levels across four aspects: accuracy, precision, recall, and F1 score, under random testing, popular testing, and adversarial testing conditions. It is important to note that larger radar charts indicate better performance.

Approach	Random				Popular				Adversarial			
	Acc$_\uparrow$	**Prec$_\uparrow$**	**Recall$_\uparrow$**	**F1$_\uparrow$**	**Acc$_\uparrow$**	**Prec$_\uparrow$**	**Recall$_\uparrow$**	**F1$_\uparrow$**	**Acc$_\uparrow$**	**Prec$_\uparrow$**	**Recall$_\uparrow$**	**F1$_\uparrow$**
Q-Sample	0.877	0.981	0.777	0.867	0.87	0.961	0.777	0.859	0.856	0.923	0.776	0.843
Q-Beam	0.884	0.964	0.805	0.877	0.877	0.940	0.805	0.867	0.859	0.902	0.805	0.851
Q-OPERA	-	-	-	-	-	-	-	-	-	-	-	-
Q-CARE	0.883	0.978	0.791	0.875	0.876	0.952	0.7913	0.864	0.859	0.915	0.791	0.849
L-Sample	0.875	0.969	0.783	0.866	0.859	0.929	0.778	0.847	0.838	0.883	0.780	0.828
L-Beam	0.884	0.960	0.810	0.878	0.860	0.903	0.807	0.852	0.826	0.840	0.805	0.822
L-OPERA	0.881	0.955	0.807	0.875	0.859	0.901	0.807	0.851	0.825	0.838	0.807	0.822
L-CARE	0.885	0.967	0.804	0.878	0.860	0.908	0.800	0.851	0.826	0.844	0.798	0.821

Table 7: POPE evaluation on hallucinations. In the Approach column, Q denotes Qwen2-VL and L denotes LLaVA-1.5.