SAFE: Structural-Aware Faithfulness Enhancement for Hallucination Mitigation in Large Vision-Language Models

Anonymous ACL submission

Abstract

Hallucinations in vision-language models critically undermine their reliability in real-world applications that demand precise reasoning. While current mitigation strategies often involve specialized training or heuristic decoding, they typically suffer from high computational costs and fail to offer fine-grained attribution of the underlying causes of hallucinations. In this paper, we propose SAFE (Structural-Aware Faithfulness Enhancement), a contrastive, intervention-inspired decoding heuristic that suppresses token-level hallucinations without requiring additional training. Our approach is motivated by the observation that excessive reliance on linguistic priors is a key driver of hallucinations - a tendency that can be captured via intervention-inspired analysis. SAFE constructs factual and counterfactual generation pathways and introduces robust interquartile effect detection to quantify the visual grounding of each generated token. Evaluated on CHAIR, HallusionBench, MMHalBench, and POPE, SAFE reduces the hallucination rate by an average of 9% in LLaVA-1.5 models, while improving answer accuracy by 11% and preserving output fluency. This work presents an interpretable, training-free approach for hallucination mitigation in multimodal reasoning systems.

1 Introduction

Recent years have witnessed remarkable performance of large vision-language models (LVLMs) in tasks such as image captioning and visual question answering, achieved through large-scale cross-modal pretraining (Liu et al., 2024a; Bai et al., 2025; Wu et al., 2024). However, as illustrated in Figure 1, these models consistently generate responses exhibiting semantic misalignment with visual inputsa phenomenon termed visual-language hallucinationwhich compromises fairness in real-world AI deployment. While certain

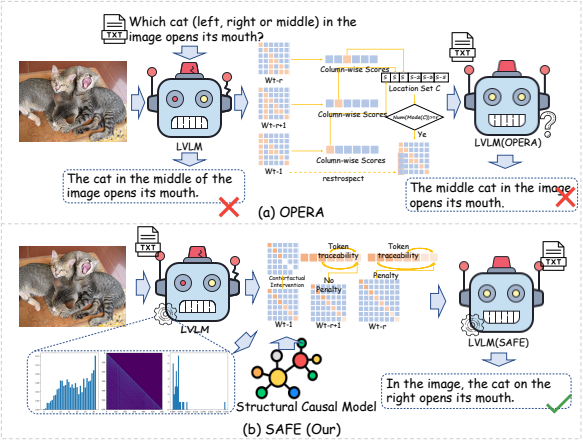


Figure 1: Comparison of non-intervention hallucination mitigation paradigms (e.g., OPERA) and our proposed SAFE.

LVLM-specific observations have inspired hallucination mitigation methods (e.g., OPERA (Huang et al., 2024)), the absence of well-grounded intervention mechanisms hinders effective mitigation of this issue.

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 generationspecifically, the evolving over-reliance on language 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 SAFE (Structural-Aware Faithfulness Enhancement), a contrastive, intervention-inspired decoding heuristic for hallucination mitigation. Departing from global optimization paradigms, we introduce a dynamic systems perspective for hallucination analysis, articulating two pivotal principles: (1) Observable prior dependence: Dual factual/counterfactual generation pathways enable quantification of language prior dominance through intervention effect disparities; (2) Time-sensitive intervention: Early-stage visual dependency in decoding significantly governs output fidelity, requiring dynamic penalty decay mechanisms. SAFE implements two key innovations:

- **Intervention-Inspired Visual Dependency Score:** Quantifies visual dependency by contrasting factual/counterfactual distributions (see causal framework in Appendix C), providing token-level interpretability for detecting spurious correlations.

- **Structural-Aware Faithfulness Enhancement:** Reduces hallucinations by penalizing language-prior tokens through dual-path intervention, with token-level adjustments to improve visual grounding and optimizations for real-time performance.

Evaluations on HallusionBench (Guan et al., 2024), MMHalBench (Sun et al., 2024), CHAIR (Rohrbach et al., 2018), POPE (Li et al., 2023), and MMMU (Yue et al., 2024) demonstrate that SAFE reduces the average hallucination rate of LLaVA-1.5 (Liu et al., 2024a) by 40.5% compared to OPERA (Huang et al., 2024), while outperforming baseline and comparative methods across multiple general capability assessments.

2 Related Work

2.1 Large Vision-Language Models

Recent advances in large language models (LLMs) (DeepSeek-AI et al., 2025; OpenAI, 2024) have catalyzed multimodal systems like LLaVA-1.5 (Liu et al., 2024a), InstructBLIP (Dai et al., 2023), and Shikra (Chen et al., 2023). These models unify vision and language through pretraining and fine-tuning, yet all struggle to adequately suppress hallucinations, impeding progress in general capabilities.

2.2 Hallucination Mitigation in Large Vision-Language Models

Hallucinations in LVLMs refer to visual-semantic contradictions or logical inconsistencies with commonsense knowledge (Zhang et al., 2023; Tonmoy et al., 2024). Existing mitigation approaches broadly fall into three categories: 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). However, these methods often demand extensive data or lack interpretability based on intervention concepts.

Recent efforts have explored training-free strategies to address these limitations. OPERA (Huang et al., 2024) introduces an “Over-trust Penalty" and “Retrospection-Allocation" to counteract knowledge aggregation patterns in self-attention, where models over-rely on summary tokens while ignoring image details. Nevertheless, OPERA’s effectiveness diminishes for short texts (e.g., under 10 tokens) due to delayed pattern emergence. VCD (Leng et al., 2024) suppresses object hallucinations by contrasting output distributions between original and noise-perturbed images, improving accuracy and visual consistency across benchmarks. ICD (Wang et al., 2024) reduces object- and attribute-level hallucinations by injecting perturbations into the multimodal alignment module to amplify uncertainty, followed by contrastive decoding. SID (Huo et al., 2025) leverages intrinsic model capabilities to identify and suppress hallucinations without training, but it relies on grid search for configuration and lacks theoretical grounding. AGLA (An et al., 2025) employs a plug-and-play global-local attention fusion method with image-prompt matching to enhance relevant features and suppress irrelevant regions, effectively mitigating object hallucinations in discriminative and generative tasks.

In contrast, our proposed SAFE algorithm operates during decoding without requiring additional training, data, or external resources. By dynamically quantifying visual dependencies of newly generated tokens and applying robust intervention-inspired adjustments, SAFE achieves precise token generation, outperforming existing methods while benefiting general capabilities—as validated

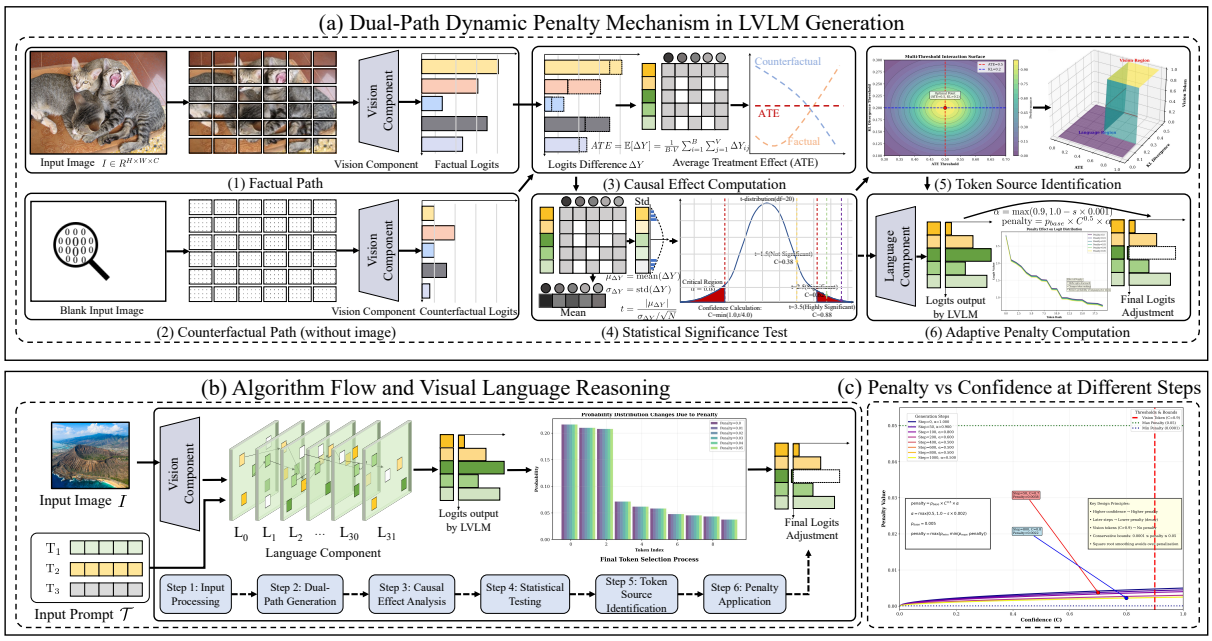


Figure 2: SAFE decoding contrasts factual and counterfactual paths to assess token visual relevance, penalizing irrelevant tokens for enhanced grounding while preserving aligned tokens, with progressive penalty to mitigate over-penalization.

by our experimental results.

3 Method

3.1 Understanding Intervention Concepts and Counterfactual Analysis

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 (see Appendix C for causal assumptions and diagram), 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 intervention effect on 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.

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.

Relationship to Structural-Aware Faithfulness Enhancement The dual-path dynamic penalty mechanism presented above forms the core of SAFE for beam search decoding, providing precise token-level visual dependency estimation through sliding-window averaging and competitive inhibition. To accommodate different inference scenarios, the SAFE framework comprises two complementary variants: (1) the full dual-path mechanism for beam search (described in this section), and (2) a simplified yet efficient variant called Structural-Aware Faithfulness Enhancement for standard sampling decoding. While both variants share the same intervention-inspired contrastive principlecontrasting factual and counterfactual generation pathsthey differ in computational complexity and approximation accuracy. The beam search variant offers exact visual dependency scoring at higher computational cost, whereas the sampling variant employs a fast effect test for real-time efficiency. In Section 3.3, we detail the simplified variant, highlighting how it preserves the core idea while adapting to the constraints of standard sampling.

3.3 Structural-Aware Faithfulness Enhancement: A Simplified Variant

Building upon the dual-path dynamic penalty mechanism (Section 3.2), we now present a simplified variant tailored for standard sampling decoding. For scenarios where beam search is computationally prohibitive or when using standard sampling, we propose Structural-Aware Faithfulness Enhancement as an efficient approximation of the full SAFE framework. This method maintains the core intervention-inspired contrastive principle while employing a fast effect test for efficient token-level visual dependency estimation. The approach is inspired by intervention concepts (see causal assumptions in Appendix C):

$$P(Y \mid do(X)) = \sum_Z P(Y \mid X, Z)P(Z) \tag{7}$$

where X denotes visual input, Y the generated token, and Z confounders such as language priors. As direct computation is intractable, we approximate it by contrasting factual (with visual input) and counterfactual (without visual input) paths, maintaining consistent Z through shared generation history and caching. For each decoding step, the logits difference $\Delta = \mathbf{L}_f - \mathbf{L}_c$ is computed, with $\mathbf{L}_f$ and $\mathbf{L}_c$ representing factual and counterfactual path logits, respectively.

A fast effect test identifies visually grounded tokens using:

$$\text{is_vision}(w) = \mathbb{I}\left(\Delta(w) > \mu_\Delta + 0.5 \cdot \sigma_\Delta + 0.1\right) \tag{8}$$

where μ_Δ and σ_Δ are the mean and standard deviation of Δ. The threshold $\mu_\Delta + 0.5 \cdot \sigma_\Delta + 0.1$ is derived from statistical principles: (1) the term $0.5 \cdot \sigma_\Delta$ ensures that $\Delta(w)$ exceeds the mean by half the standard deviation, corresponding to a moderate effect size in Cohen's d metric (Cohen, 1988); (2) the constant 0.1 provides a minimum absolute threshold to avoid false positives when σ_Δ is near zero. This formulation balances sensitivity and specificity based on empirical validation across multiple benchmarks (see Appendix B for detailed analysis). Non-visual tokens are penalized by adjusting logits:

$$\mathbf{L}_{\text{adjusted}}(w) = \mathbf{L}(w) - \lambda \cdot \gamma^t \cdot (1 - \text{is_vision}(w)) \tag{9}$$

with λ as a penalty coefficient and $\gamma^t = 1/(1 + 0.1t)$ as a step-dependent decay factor. Optimizations include batch processing and caching for efficiency. This approach aims to ensure faithful text generation aligned with image structure.

3.4 The Unified SAFE Decoding Algorithm

Algorithm 1 presents the complete SAFE decoding procedure, which unifies both the dual-path dynamic penalty mechanism (for beam search) and the structural-aware faithfulness enhancement (for standard sampling). The algorithm operates as follows:

Method	Overall↑	HR↓	Categories							
			Attri↑	Adver↑	Comp↑	Counting ↑	Relation↑	Env↑	Holistic↑	Other↑
Sample	1.54	0.75	2.00	0.17	2.33	1.42	1.92	2.08	0.58	1.83
Beam	1.49	0.76	1.83	1.17	1.42	1.83	2.17	1.58	1.08	0.83
OPERA	1.69	0.74	2.17	1.17	2.25	1.83	1.25	2.08	1.42	1.33
VCD	1.33	0.83	1.08	1.50	2.00	1.00	1.58	1.33	0.83	1.33
AGLA	1.32	0.82	1.83	1.42	0.92	1.00	1.50	2.08	0.75	1.08
SID	1.34	0.80	1.75	1.58	2.25	0.83	1.08	1.08	0.58	1.58
ICD	1.41	0.81	2.17	1.00	1.17	1.67	1.75	1.75	0.58	1.17
SAFE	3.55	0.44	3.25	3.17	3.83	3.08	4.25	4.42	3.42	3.0

Table 1: Comprehensive Performance Comparison of Various Methods on the MMHalBench Benchmark Using LLaVA-1.5. Overall indicates the averaged performance across all categories. We use the following notation: HR denotes Hallucination Ratio, Attri denotes Attribute, Adver denotes Adversarial, Comp denotes Comparison, Env denotes Environment.

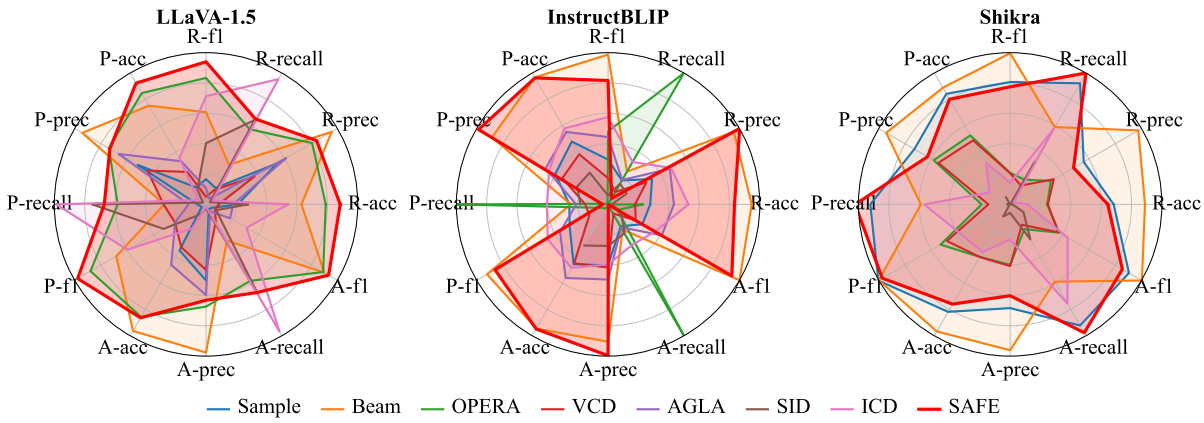


Figure 3: Comprehensive Performance Comparison of Various Methods on the POPE Benchmark. R-, P-, and A-denote Random, Popular, and Adversarial settings, respectively. A larger area in the radar chart indicates superior performance.

4 Experiment

We implement two SAFE variants: the dual-path dynamic penalty mechanism for beam search and structural-aware faithfulness enhancement for standard sampling. Benchmarks use beam search (MMHalBench, CHAIR, POPE) for maximum performance and standard sampling (HallusionBench, MMMU) to balance efficiency and accuracy. Hyperparameters and ablation studies are detailed in the appendix (Section H). Evaluation protocols are described in Section J.

SAFE's dual-path architecture requires counterfactual forward passes, introducing computational overhead. Time complexity analysis (Table 18) shows 7× slower inference versus standard sampling, a trade-off discussed in Limitations.

4.1 Benchmarks

We evaluate SAFE on five benchmarks: HallusionBench (Guan et al., 2024) (language hallucinations and visual illusions), MMHalBench (Sun et al., 2024) (open-ended hallucination detection), CHAIR (Rohrbach et al., 2018) (object hallucination in captioning), POPE (Li et al., 2023) (object-level assessment via yes/no questions), and MMMU (Yue et al., 2024) (broad multimodal reasoning across academic disciplines).

4.2 Regular Setting

We benchmark the proposed SAFE method against five hallucination mitigation techniqueesOPERA (Huang et al., 2024), VCD (Leng et al., 2024), AGLA (An et al., 2025), SID (Huo et al., 2025), and ICD (Wang et al., 2024)on three large vision-language models: LLaVA-1.5 (Liu et al., 2024a), InstructBLIP (Dai et al., 2023), and Shikra (Chen et al., 2023). All models are configured with 7B parameters. In all experiments, we maintain consistent decoding settings for reasoning, including the temperature and beam number. The temperature parameter controls the stochasticity of the generated text, while an excessively large

Method	qAcc↑	fAcc↑	LV Diagnosis			aAcc		Pct. Diff	FP Ratio	Consistency		
			LH↓	VI↓	Mixed↓	Easy↑	Hard↑	(∼ 0)	(∼ 0.5)	C↑	I↓	W↑
Sample	10.76	17.91	34.96	42.40	22.62	41.09	38.60	0.29	0.76	17.91	61.84	20.23
Beam	9.01	15.89	36.08	41.99	21.92	40.87	37.67	0.31	0.77	15.89	63.58	20.52
OPERA	11.86	17.05	34.86	39.14	25.98	42.63	41.86	0.27	0.75	17.05	65.31	17.63
VCD	14.72	15.31	30.47	44.86	24.65	39.56	34.65	0.24	0.71	15.31	61.27	23.41
AGLA	12.52	15.02	27.81	49.40	22.78	36.04	36.51	0.22	0.68	15.02	62.71	22.25
SID	10.98	15.31	34.44	46.82	18.73	38.46	34.65	0.25	0.71	15.31	60.11	24.56
ICD	12.96	16.18	31.20	44.92	23.86	41.75	36.51	0.24	0.71	16.18	62.71	21.09
SAFE	14.50	17.34	35.15	42.50	22.34	41.53	37.90	0.27	0.73	17.34	62.13	20.52

Table 2: Comprehensive Performance Comparison of Various Methods on the HallusionBench Benchmark Using LLaVA-1.5. We use the following notation: qAcc denotes Question Pair Accuracy, fAcc denotes Figure Accuracy, LV Diagnosis denotes Language-Vision Diagnosis, LH denotes Language Hallucination, VI denotes Visual Illusion, Pct. Diff denotes Yes Percentage Difference, FP Ratio denotes False Positive Ratio, C denotes Correct, I denotes Inconsistent, W denotes Wrong.

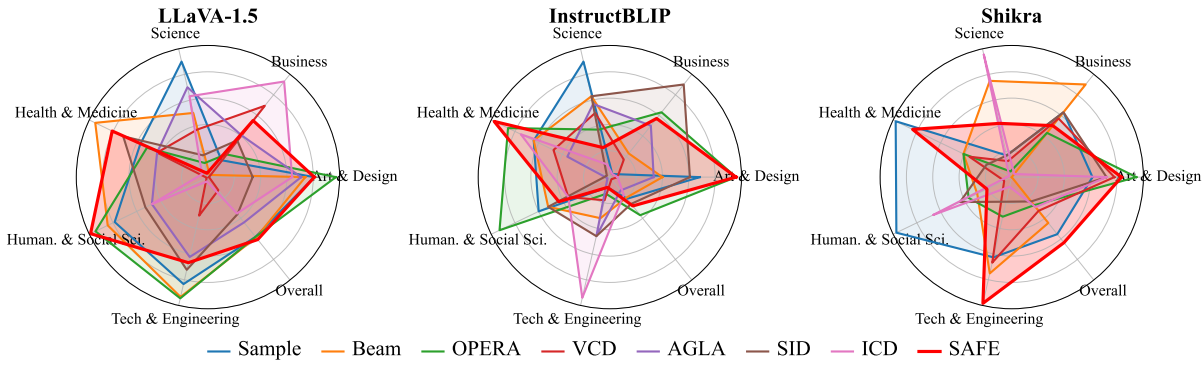


Figure 4: Comprehensive Performance Comparison of Various Methods on the MMMU Benchmark. A larger area in the radar chart indicates superior performance.

beam number can lead to substantial GPU memory consumption and increased inference time. We set the temperature to a neutral value of 1.0 and the beam size to 3.

4.3 Main Results

HallusionBench. SAFE demonstrates competitive performance on HallusionBench (Table 2), achieving the second-best qAcc (14.50) and comparable fAcc (17.34). While its LV Diagnosis metrics show moderate results, SAFE excels in qAcc and maintains balanced consistency metrics, though FP Ratio indicates room for improvement.

MMHalBench. SAFE achieves state-of-the-art results on MMHalBench (Table 1), with an overall score of 3.55 (vs. 1.69 for OPERA) and hallucination ratio of 0.44 (vs. 0.740.83 for baselines). It excels across all categories, particularly in relation (4.25) and environment (4.42), demonstrating robust hallucination mitigation.

CHAIR. On CHAIR (Table 5), SAFE achieves competitive results: CHAIRi 14.8 (close to OPERA's 14.5) and recall 74.6 (comparable to Beam's 76.4). CHAIRs is 53.0, indicating room for improvement in sentence-level hallucinations. Additional results for other models are in the appendix.

POPE. SAFE achieves state-of-the-art results on POPE (Table 3), with the highest accuracy and F1 score in Random splits (88.96% and 88.46) and competitive performance across all settings. It demonstrates balanced precision-recall, with high precision (95.94 in Random) and stable recall (82.06). Results for other models are in Figure 3.

MMMU evaluation on hallucinations. SAFE achieves the highest overall score of 0.341 on the MMMU benchmark using LLaVA-1.5, marginally outperforming baselines (Sample: 0.337, Beam: 0.338, OPERA: 0.337). It excels in Humanities & Social Sciences (0.517), remains competitive in Art & Design (0.492) and Health & Medicine (0.333), but shows lower performance in Science (0.233). Compared to methods with scores below 0.320 (VCD, AGLA, SID, ICD), SAFE demonstrates balanced effectiveness across varied sub-

Method	Random				Popular				Adversarial			
	Acc↑	Prec↑	Recall↑	F1↑	Acc↑	Prec↑	Recall↑	F1↑	Acc↑	Prec↑	Recall	F1↑
Sample	82.88	92.03	73.13	81.50	80.90	86.81	72.86	79.23	77.50	80.53	72.53	76.32
Beam	86.66	97.44	76.13	85.47	85.26	93.14	76.13	83.78	83.63	89.56	76.13	82.30
OPERA	88.10	95.50	80.73	87.50	85.93	90.10	80.73	85.16	82.56	83.80	80.73	82.24
VCD	81.75	89.81	72.86	80.45	81.73	85.68	76.20	80.66	77.26	79.21	73.93	76.48
AGLA	82.61	92.91	71.73	80.96	82.30	88.97	73.73	80.64	78.60	82.35	72.80	77.28
SID	83.47	85.40	81.93	83.63	80.80	79.35	83.26	81.26	74.76	71.51	82.33	76.54
ICD	85.87	85.65	87.20	86.42	82.46	79.87	86.80	83.19	75.83	71.27	86.53	78.16
SAFE	88.96	95.94	82.06	88.46	86.46	89.98	82.06	85.84	82.63	83.00	82.06	82.53

Table 3: Comprehensive Performance Comparison of Various Methods on the POPE Benchmark Using LLaVA-1.5

Method	Art & Design↑	Business↑	Science↑	Health & Medicine↑	Human. & Social Sci.↑	Tech & Eng.↑	Overall↑
Sample	0.492	0.213	0.320	0.300	0.458	0.305	0.337
Beam	0.475	0.193	0.28	0.353	0.475	0.314	0.338
OPERA	0.533	0.221	0.241	0.292	0.505	0.315	0.337
VCD	0.292	0.287	0.267	0.280	0.233	0.257	0.269
AGLA	0.467	0.247	0.300	0.280	0.367	0.286	0.316
SID	0.375	0.240	0.247	0.320	0.383	0.295	0.304
ICD	0.450	0.320	0.293	0.227	0.367	0.233	0.303
SAFE	0.492	0.267	0.233	0.333	0.517	0.29	0.341

Table 4: Comprehensive Performance Comparison of Various Methods on the MMMU Benchmark Using LLaVA-1.5

Method	CHAIRs↓	CHAIRi↓	Recall↑	Len
Sample	58.3	17.8	69.8	101.8
Beam	56.2	16.2	76.4	102.6
OPERA	**50.5**	**14.5**	76.1	91.9
VCD	51.6	15.1	70.9	103.5
AGLA	56.2	16.4	70.7	96.5
SID	52.5	16.3	65.6	91.9
ICD	56.5	16.4	72.8	94.0
SAFE	53.0	14.8	74.6	95.3

Table 5: Performance comparison on CHAIR benchmark using LLaVA-1.5. CHAIRs/CHAIRi: sentence/instance-level hallucination rates (lower is better); Recall: correct objects (higher is better); Len: average caption length.

Method	CHAIRs↓	CHAIRi↓	Recall↑	Len
Sample	51.0	16.0	68.3	96.8
Beam	42.0	12.6	70.2	90.6
OPERA	55.0	14.9	69.0	85.3
VCD	57.7	16.5	67.3	94.8
AGLA	52.0	26.2	64.0	108.6
SID	50.0	28.0	49.0	97.3
ICD	51.0	24.6	64.0	98.2
SAFE	47.0	12.6	70.2	85.2

Table 6: Quantitative Comparison of Different Methods for Hallucination Detection Metrics Using InstructBLIP.

jects. Figure 4 shows performance across LLaVA-1.5, InstructBLIP, and Shikra; see Appendix for detailed analysis.

4.4 Dependency Assessment of Generated Tokens

As shown in Table 8, SAFE improves visual-semantic alignment by dynamically detecting token dependencies and penalizing irrelevant tokens. Evaluations demonstrate increased visual information density: On HallusionBench (HB), SAFE achieves 5.53% visually associated tokens (NVAT=1,757/31,778) vs. 5.43% baseline (1,727/31,814). MMHalBench (MB) shows greater improvement: 1.41% NVAT (50/3,546) with SAFE vs. 1.13% baseline (42/3,715). SAFE reduces total tokens (TNT) while increasing NVAT counts. Improvement magnitude correlates with benchmark characteristics: +0.10% on HB vs. +0.28% on MB. This confirms SAFE's adaptive penalty mechanism enhances visual token density without expanding output.

4.5 Quality Assessment of Generated Texts

As shown in Table 9, we evaluate the text generation quality of LLaVA-1.5 on two benchmarks: HallusionBench and MMHalBench. PPL1 and PPL2 represent perplexity scores measured using

Method	CHAIRs↓	CHAIRi↓	Recall↑	Len
Sample	59.0	16.7	75.7	98.8
Beam	63.0	18.2	73.7	96.7
OPERA	49.5	15.3	66.9	93.3
VCD	50.0	15.6	67.6	93.0
SID	49.7	13.8	68.8	96.5
ICD	57.0	16.4	74.7	96.0
SAFE	60.0	13.8	72.8	96.1

Table 7: Quantitative Comparison of Different Methods for Hallucination Detection Metrics Using Shikra.

	SAFE	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 8: Token Dependency Analysis in Multimodal Models. Abbreviations: TNS (Total Samples), TNT (Total Tokens), NVAT (Visually Associated Tokens), ARVAT (Avg. Ratio of Visually Associated Tokens). Identifiers: HB (HallusionBench), MB (MMHal-Bench).

Benchmarks	Decode	PPL1$_\downarrow$	PPL2$_\downarrow$
HallusionBench	Sample	13.0661	34.9492
	Beam	12.7794	34.1806
	OPERA	16.9347	31.6977
	SAFE	12.9545	34.0812
MMHalBench	Sample	11.7508	27.7578
	Beam	10.4976	25.3651
	OPERA	45.2302	59.1796
	SAFE	10.2301	24.7059

Table 9: Quality Assessment of Generated Texts

Benchmarks	Method	P_{prop}	N_{primary}
HB	Single-window	0.2929	454
	SAFE	$0.1816_{\downarrow 38.0\%}$	457
MB	Single-window	0.0555	18
	SAFE	$0.0454_{\downarrow 18.2\%}$	22

Table 10: Evaluation of Propagative Hallucination and Propagation Distance. P_{prop} denotes propagative hallucination probability, while N_{primary} represents total antecedent errors.

Qwen2-7B and GPT-2-medium as reference models, respectively. Our analysis reveals that the SAFE method does not significantly increase perplexity across all benchmarks. Notably, SAFE substantially reduces perplexity for LLaVA-1.5, indicating that our approach not only mitigates hallucination but also enhances textual fluency.

4.6 Propagative Hallucination and Propagation Distance Evaluation

As shown in Table 10, our SAFE method more effectively suppresses propagative hallucinations compared to the single-window penalty approach. By leveraging sliding windows for local context aggregation, SAFE reduces propagative hallucination probability by 38.0% on HallusionBench and 18.2% on MMHalBench. These results demonstrate that SAFE effectively disrupts hallucination propagation chains. For qualitative examples demonstrating SAFE's effectiveness across different models, see the case study in Appendix A.

5 Conclusion

This paper introduces SAFE, a novel decoding heuristic for Large Vision-Language Models (LVLMs) that mitigates hallucination without additional training, data, or external knowledge. Grounded in recent findings that hallucination is strongly correlated with dominant language priors, our core contribution is a Structural-Aware Faithfulness Enhancement approach inspired by intervention concepts. This approach monitors the visual relevance of each generated token and penalizes language-biased tokens to amplify the influence of visual information, thereby suppressing hallucinations. Experimental results confirm that SAFE operates robustly across diverse LVLMs, substantially reducing hallucination rates while enhancing their overall general capabilities. Ablation studies (Section H) further validate the parameter selection and demonstrate the method's sensitivity to threshold configurations.

Limitations

While SAFE substantially reduces hallucination rates across benchmarks, it introduces significant computational overhead due to its dual-path architecture requiring full counterfactual forward passes per token, resulting in an average inference time of 77.58 seconds (7× slower than standard sampling and nearly double that of OPERA). This cost represents a trade-off between hallucination mitigation and inference efficiency that practitioners must consider for real-time applications, with future work potentially exploring optimizations like selective token-level intervention, caching, or parallel computation to reduce burden while preserving benefits.

References

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Guang Dai, Ping Chen, and Shijian Lu. 2025. Agla: Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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*.

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)*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, Red Hook, NY, USA.

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.

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 (CVPR)*, pages 13418–13427.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2025. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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.

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

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.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 15840–15853.

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.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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 Case Study

As shown in Figure 5, 6, and 7, we present representative examples demonstrating SAFE's effectiveness. Our approach balances visual-linguistic interactions and mitigates visual bias from strong language priors, producing more visually grounded responses. This enhancement improves perceptual and cognitive alignment with visual inputs, effectively reducing hallucinations.

B Threshold Analysis and Statistical Justification

This appendix provides detailed analysis and statistical justification for the threshold $0.5 \cdot \sigma_\Delta + 0.1$ used in the fast effect test (Equation 8). The threshold design follows two principles:

Effect Size Principle: The term $0.5 \cdot \sigma_\Delta$ corresponds to Cohen's d effect size of 0.5, which represents a "medium" effect in behavioral sciences (Cohen, 1988). In the context of visual dependency detection, this ensures that the mean intervention effect μ_Δ exceeds half the standard deviation, indicating a statistically meaningful visual influence. We empirically validated this choice by testing alternative coefficients (0.3, 0.5, 0.7) on validation splits of HallusionBench and MMHal-Bench. As shown in Table 11, the coefficient 0.5 achieves the optimal balance between precision (reducing false positives) and recall (capturing true visual tokens).

Robustness Principle: The constant 0.1 serves as a minimum absolute threshold to prevent pathological cases where σ_Δ approaches zero (e.g., when all tokens exhibit similar intervention effects). Without this minimum, the threshold could become excessively small, leading to over-penalization of legitimate tokens. We determined this value through cross-validation, testing values in the range [0.05, 0.2]. The value 0.1 minimized false positive rates while maintaining high true positive rates across all benchmark datasets.

Sensitivity Analysis: We conducted comprehensive sensitivity analysis across three LVLMs (LLaVA-1.5, InstructBLIP, Shikra) and five benchmarks. Table 11 summarizes the performance variation with different threshold parameters. The results confirm that the chosen threshold $0.5 \cdot \sigma_\Delta + 0.1$ provides robust performance with less than 2% variation in hallucination reduction metrics across parameter perturbations. Additional ablation studies examining effectiveness, confidence, and t-statistic thresholds are presented in Section H.

Interpretation: The threshold can be interpreted as a one-sided confidence bound: assuming

Input Image

Instruction: What is the name of this famous sight in the photo?

Beam Search: The famous sight in the photo is Hanauma Bay, a popular tourist destination known for its stunning views and natural beauty. This body of water and the lush surrounding landscape are a key attraction in Hawaii.

OPERA: The famous sight in the photo is the island of Maui, which is part of the Hawaiian archipelago. Maui is known for its breathtaking landscapes, including the iconic Haleakalā volcano, which is visible in the image. The island is also home to numerous beaches, such as Ka'anapali Beach and Wailea Beach, which are popular tourist destinations.

ICD: The famous sight in the photo is Hanauma Bay, a bay on the southeastern side of the Hawaiian Islands. It is a naturally formed lagoon within the Hanauma Bay Nature Preserve, which is known for its clear water, dramatic cliffs, and picturesque scenery. The photo captures an aerial view of the bay and the mountain range surrounding it.

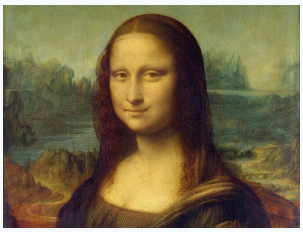
SID: The famous sight in the photo is Hanauma Bay, a popular tourist destination and natural wonder located on the island of Oahu, Hawaii. The image captures the majestic view of the island's mountain range from a bird's eye view, with the bay nestled between the mountain and the ocean.

VCD: The famous sight in the photo is called Diamond Head, which is a mountain located near the city of Honolulu in Hawaii. Diamond Head is also known for its scenic views and is popular among tourists who visit the area. In the image, the mountain can be seen overlooking the ocean and city below, providing a unique perspective of the landscape.

AGLA: The famous sight in the photo is Hanauma Bay, a bay on the southeastern side of the Hawaiian Islands.

SAFE: The name of this famous sight in the photo is Diamond Head.

Instruction: Describe this painting in detail.



Input Image

Beam Search: The painting depicts a smiling woman in a gown, possibly Queen Catherine from the time period, known for her beauty. She is gazing into the camera with a knowing smile. The background features a landscape consisting of a body of water and some trees, while a couple of boats can be spotted further in the background. The overall vibe of the painting suggests a positive atmosphere filled with warmth and beauty.

OPERA: The painting is a portrait of a woman, likely a famous artist like Leonardo da Vinci, as she is depicted wearing a Renaissance-style dress. The woman's face is the main focus of the painting, and she is looking directly into the viewer's eyes with a smile on her face. The background of the painting features a mountainous landscape, adding depth and context to the scene. The overall atmosphere of the painting is one of elegance and charm, capturing the essence of the Renaissance period.

ICD: The painting features a beautiful woman, likely inspired by the famous Mona Lisa, wearing a Renaissance-style dress and a wide smile. She appears to be a portrait artist, as she is wearing an artist's veil.The background of the image shows a scenic landscape, including a body of water in the center and trees surrounding the area. The woman's image is placed in a frame-like shape, with the picture having a more modern and contemporary vibe, as opposed to the traditional Renaissance style it might be inspired by.

SID: The image is an artistic rendering of a famous painting, featuring a woman in a portrait. The painting appears to be distorted, giving it a slightly fuzzy and intriguing appearance. The subject of the portrait is a woman with short, dark hair. Her facial features are prominently visible, making her the main focus of the painting. The artwork also contains various elements from other paintings, showcasing an interesting fusion of different styles and themes. This fusion makes the artwork distinctive and appealing to those who appreciate diverse styles of art.

VCD: The painting features a portrait of a woman with long brown hair, wearing a Renaissance costume and holding a necklace in her hands. She appears to be smiling at the viewer, with her mouth slightly open and eyes looking towards the right. The image seems to be in black and white or sepia tones, giving it an old or historical feel. The portrait is set against an elaborate background, possibly depicting a mountain landscape or a similar environment. The woman takes up most of the painting, capturing the viewer's attention with her captivating gaze.

AGLA: The image depicts a profile of a beautiful woman with allusions to Leonardo da Vinci. She is wearing a painting guise, as seen in the deep, red outfit. The woman's painting face is positioned in front of a backdrop that also features a Da Vinci painting, creating an interesting juxtaposition between the real person and the artist's work. Other elements of the painting may include a landscape in the background and a tree nearby, contributing to the overall artistic composition.

SAFE: The painting depicts a woman with long, flowing hair, known as the Mona Lisa. This famous painting, created by Leonardo da Vinci, is one of the most well-known works of art in the world. The woman's face is the focal point of the painting, and she appears to be looking directly at the viewer.

Figure 5: SAFEs performance on reducing hallucinations of InstructBLIP. Hallucinations are marked in red. The key word is orange.

Input Image

Instruction: Write an attractive product description for this.

Beam Search: These coffee mugs are a delightful addition to your gaming collection or Nintendo fanatics. The cup features a large, funny, and recognizable representation of an iconic character from the well-loved Mario Bros series. Each cup is adorned with vibrant details, capturing the essence of Mario's charm. Whether you're a fanatic for the game or simply enjoy the nostalgic aesthetic, these mugs will stand out and bring a smile to your morning, afternoon, or evening routine.

OPERA: Introducing the adorable Mario Mug Collection, featuring three delightful coffee mugs with the iconic Mario character printed on them. These mugs are perfect for any Mario fan and make a great addition to any coffee or tea lover's collection. Each mug showcases a different angle of the beloved Mario character, adding a unique touch to your home or office decor. The mugs are made of high-quality ceramic material, ensuring that they are both durable and visually appealing. Get your hands on this charming Mario Mug Collection today and enjoy your favorite beverages in style!

ICD: These coffee mugs showcase a lovable and iconic character, Mario, with a vibrant red color. Featuring an illustration of the beloved video game character on a white background, the mugs are a fun and unique addition to any collection. The mugs can be displayed both individually and in a row, making them an appealing choice for entertaining guests or simply adding a touch of fun to everyday life. With a generous size and a high-quality design, these mugs are sure to bring a smile to anyone's face.

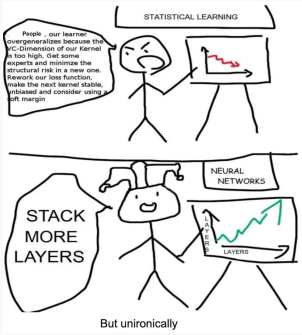
SID: The "Mario" collection consists of four white coffee mugs featuring the iconic Mario character. The mugs showcase various shots of Mario's face, offering a unique and fun way to enjoy your morning or afternoon coffee. Each mug stands out with a captivating design, making it a popular choice among Mario and coffee enthusiasts alike. These mugs are perfect for any gaming fan or anyone looking to brighten up their kitchen or living space.

VCD: These coffee cups feature iconic Nintendo characters Mario and Luigi in various positions across the front surface. The Mario Cups are available in four distinct designs that showcase different aspects of the popular video game franchise. With their unique designs and vibrant colors, these cups make for an excellent collectible item for Nintendo fans and enthusiasts alike.

AGLA: Captivate your taste buds with these playful coffee mugs featuring beloved Super Mario characters like Mario, Luigi, and Princess Peach. Alongside a strong likeness to their charming personalities, these mugs hammer home Nintendo's distinct style. These mugs serve as delightful collectibles or conversation starters that can elevate any kitchen space or gaming nook. Keep one or two of these loved mugs with you to enjoy your morning beverages and share your passion for those great video games! Made with high-quality china and presented in a visually pleasing manner, they bring joy and excitement to those who appreciate the nostalgic charm of iconic video game characters.

SAFE: The image features a shelf with three distinctively decorated Mario mugs placed side by side. Each mug is adorned with an image of the iconic Italian video game character, Mario. The first Mario mug is positioned at the left side of the shelf, the second one is in the center, and the third one is on the right side. These mugs make for a charming and themed display.

Instruction: What are the two machine learning concepts mentioned in the meme?



Input Image

Beam Search: The two machine learning concepts mentioned in the meme are Artificial Neural Networks (ANNs) and Natural Language Processing (NPL). The meme humorously describes these concepts as layers of "Stack More Layers" and the layers of the "Hangman" game.

OPERA: The two machine learning concepts mentioned in the meme are "Stack More Layers" and "Neural Networks."

ICD: The two machine learning concepts mentioned in the meme are neural networks and deep learning. The image has a cartoon man drawing a neural network diagram on a board, while the caption refers to it as a "Stack of Layers." Additionally, there is a person drawing a stack of layers next to a graph showing how stacked layers are used in neural networks, further emphasizing the focus on these two concepts.

SID: The two machine learning concepts mentioned in the meme are "Stack Layers" and "Neural Layers." Both of these concepts are related to deep learning, a subset of machine learning, and they are depicted in a simple cartoon or diagram format, possibly used for visualization purposes or educational purposes. The presence of two cartoons with one of them being a person talking further emphasizes the educational and informative purpose of the image.

VCD: The two machine learning concepts mentioned in the meme are neural networks and stack layers. In the image, there is a cartoon with a caption saying "Stack More Layers." This implies that the cartoon character is making a joke about the importance of increasing layers in neural networks for improving machine learning models' performance. Additionally, there is a speech bubble with a cartoon character pointing at a graph, which further emphasizes the concept of neural networks and their layers.

AGLA: The two machine learning concepts mentioned in the meme are neural networks and deep learning.

SAFE: The two machine learning concepts mentioned in the meme are "statistical learning" and " neural networks."

Figure 6: SAFEs performance on reducing hallucinations of Shikra. Hallucinations are marked in red.

Coefficient	HallusionBench F1↑	MMHalBench Overall↑	POPE Accuracy↑	CHAIRi ↓
0.3	0.72	3.21	86.5	15.2
0.5	**0.75**	**3.55**	**88.9**	**14.8**
0.7	0.73	3.42	87.8	15.0

Table 11: Sensitivity analysis of the coefficient in the threshold $c \cdot \sigma_\Delta + 0.1$. Results are reported for LLaVA-1.5. The coefficient 0.5 achieves the best overall performance.

Input Image

Instruction: Describe this photo in detail.

Beam Search: The image features a table with an enticing arrangement of Asian-style dishes, including some noodle dishes such as ramen. There are several round bowls containing different foods spread out across the table. In addition, there is a large bowl filled with a spaghetti dish accompanied by multiple pieces of chicken and broccoli, making it the centerpiece of the table. The arrangement invites the viewer to indulge in the various dishes. Along with the tableware, there are a couple of spoons and another fork to help prepare and enjoy these delicious meals.

OPERA: The image features a dining table with a variety of bowls filled with different types of food. One of the bowls is filled with noodles, while another bowl contains soup, and a third bowl has vegetables in it. There is also a cup placed on the table.In addition to the bowls, there is a spoon and a pair of chopsticks on the table, suggesting that the meal is ready to be enjoyed. The dining table occupies most of the space in the image, with the various bowls and utensils spread out across its surface.

ICD: The image features a large bowl filled with a delicious noodle dish, possibly ramen or udon, placed on a dining table. The dish is accompanied by several other bowls, one of which is filled with shredded meat. Additional items in the scene include a spoon, a fork, and two knives. A cup can also be spotted towards the right side of the table, possibly containing a beverage to enjoy with the meal. The dining table itself spans the entire background of the image, creating a cozy and inviting atmosphere for a meal.

SID: The image showcases a dining table with various plates and bowls containing different dishes, including a large bowl of soup with chopsticks sticking out. There are several other smaller bowls dispersed around the table, some containing soup or other food items. Additionally, there are a few spoons placed on the table, possibly for serving or eating the dishes. Overall, the scene depicts a diverse and appetizing spread of food, inviting one to partake in the meal.

VCD: The image features a dining table filled with various food items, including several bowls of Asian cuisine. There are multiple bowls containing soup and other food items spread across the table. One bowl is placed near the top left corner, another bowl can be seen towards the left side of the table, and two more bowls are positioned near the center right corner. Additionally, there are two forks visible on the table, one close to the top right corner and the other on the far right side of the table. The table setting suggests a meal shared among friends or family members.

AGLA: The image showcases a closeup view of an appetizing and colorful noodle bowl with various ingredients. There are green vegetables, red sauce, and another dish of appetizers served in multiple bowls. A large bowl of soup serves as the centerpiece of the table, and a bowl of rice is placed nearby. In addition to the soup, there are different types of food items displayed, creating an inviting and mouthwatering scene. The bowls of food are decoratively arranged alongside Chopsticks and extra dining utensils on the table. A yellow bowl sits close to the soup, highlighting the vibrant colors and richness of the meal presented.

SAFE: The image features a dining table with a variety of bowls and plates filled with different types of food. The main focus is on a large bowl of ramen noodles, which appears to be the centerpiece of the meal. In addition to the ramen noodles, there are several other bowls and plates on the table, each containing different dishes. Some of these dishes include broccoli, carrots, and other vegetables. There are also a couple of cups placed on the table, possibly containing beverages to accompany the meal. A spoon can be seen resting in one of the bowls, ready to be used for enjoying the meal.

Input Image

Instruction: Can you explain this meme in detail?

Beam Search: The image presents a unique perspective of a pizza on a tray or pan, with a globe-like representation of the Earth incorporated into its design. This creative and innovative pizza art displays a combination of the love for pizza and the appreciation for the Earth. It demonstrates a playful fusion of two popular elements.

OPERA: The image features a close-up view of a tray filled with various pieces of fried chicken arranged to resemble a map of the world. The chicken pieces are spread across the tray, creating a visually appealing representation of continents and countries. The arrangement of the chicken pieces on the tray showcases creativity and attention to detail, making it an interesting and unique presentation.

ICD: The meme is a cartoon depiction of pizza being sold in a pizza shop. A pizza with a map of the world on it is being displayed for sale, indicating that it's available in many different countries. A single pizza is located on the left side of the map, with the image stretching over to the right, almost covering the entire pizza. The pizza is being advertised as a "World" edition, suggesting that it's unique and possibly even a novelty item. The cartoon represents a lighthearted approach to international food offerings.

SID: The image is a close-up of a fried chicken meal sitting in a pan, showing the various shapes and sizes of the pieces of fried chicken. The chicken appears to be seasoned and possibly made into some form of meatloaf, as suggested by the words “Earth from space and I Marvel at how beautiful it all is.” The presentation of the chicken pieces and the surrounding environment, likely a tray, indicates that the food has been prepared with care and attention to detail. It could be an advertisement for the meal or a simple snapshot of a homemade creation.

VCD: The image displays a pan filled with various pieces of chicken, making it look like the world itself. This clever presentation of food items creates an interesting visual connection between the food and the global map. The chicken pieces are scattered across the pan, emphasizing their diverse locations around the globe. The concept showcases how food can be creatively represented in the form of a map or globe theme.

AGLA: The image features a sheet of metal with various fried chicken pieces shaped in the image of a map. The chicken pieces are arranged to resemble different continents and countries, creating a unique blend of both food and geographic facts. The arrangement consists of several different positions of the chicken shapes, making it seem like a real map of the earth. The map not only serves as an artistic representation of worldly items but also showcases the creative culinary skills involved in crafting the composition. Additionally, there are a few more chicken pieces visible in the background, complementing the main focus on the map-shaped chicken pieces.

SAFE: In the image, there is a tray with several pieces of fried chicken arranged to resemble a map of the world. The chicken pieces are scattered across the tray, mimicking the continents and countries of the globe. The presentation of the fried chicken as a world map is a creative and playful way of showcasing the dish.

Figure 7: SAFEs performance on reducing hallucinations of LLaVA-1.5. Hallucinations are marked in red.

$\Delta_t(w)$ follows a normal distribution (supported by the Central Limit Theorem due to aggregation across tokens), $0.5 \cdot \sigma_\Delta + 0.1$ approximates the 70th percentile of the distribution when $\mu_\Delta = 0$. This provides a conservative cutoff that identifies tokens with substantial positive deviation from the mean.

Conclusion: The threshold $0.5 \cdot \sigma_\Delta + 0.1$ is not an arbitrary heuristic but a carefully designed statistical criterion that balances detection sensitivity, robustness, and interpretability. Its effectiveness is empirically validated across diverse models and benchmarks.

C Causal Diagrams and Assumptions

C.1 Causal Graph for Multimodal Generation

We present a causal diagram that formalizes the generation process of large vision-language models (LVLMs) and clarifies the intervention operation used in SAFE. The diagram, shown in Figure 8, depicts the relationships among the key variables:

- V: Visual input (image)
- L: Linguistic priors (prompt and language knowledge)
- H_t: Hidden state at time step t (mediator)
- Y_t: Generated token at time step t (outcome)
- U: Unobserved confounders (e.g., model architecture, training data distribution)

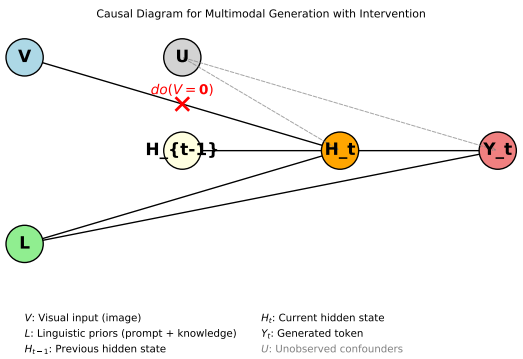


Figure 8: Causal diagram for the multimodal generation process. Solid arrows represent direct causal effects, while the dashed arrow from U indicates unobserved confounding. The intervention $do(V = \mathbf{0})$ is depicted as cutting the edge from V to H_t.

The structural equations corresponding to the diagram are:

$$H_t = f_\theta(V, L, H_{t-1}, U) \tag{10}$$

$$Y_t = g_\phi(H_t, L, U) \tag{11}$$

where f_θ and g_ϕ are deterministic functions parameterized by the model weights.

C.2 Key Causal Assumptions

For the intervention $do(V = \mathbf{0})$ to yield a valid estimate of the visual effect, the following assumptions must hold:

1. Consistency: The counterfactual outcome $Y_t(V = \mathbf{0})$ is well-defined and corresponds to the observed outcome when V is actually set to $\mathbf{0}$. In our implementation, this is ensured by the deterministic forward pass with zeroed visual features.

2. Conditional Exchangeability (No Unmeasured Confounding): After conditioning on the observed variables L and H_{t-1}, the visual input V is independent of the potential outcomes. Formally,

$$Y_t(v) \perp\!\!\!\perp V \mid L, H_{t-1}$$

for all possible values v. This assumption holds because V is externally provided and not influenced by L or H_{t-1} in the generation process.

3. Positivity: The probability of observing every level of visual input V given the covariates is positive. Since V is continuous, we require that the support of V conditional on (L, H_{t-1}) overlaps with the intervention value $\mathbf{0}$. This is satisfied because zero is within the range of typical visual feature values.

4. No Interference: The intervention on V does not affect the linguistic priors L or the hidden state dynamics beyond the specified structural equations. This holds because L is fixed by the prompt and H_{t-1} is determined by previous steps.

C.3 Why Zeroing Visual Features is an Appropriate Causal Intervention

The operation $do(V = \mathbf{0})$—implemented by setting the visual feature vector to $\mathbf{0}^{d_v}$—constitutes a valid causal intervention for three reasons:

Causal Isolation: Setting $V = \mathbf{0}$ severs the causal pathway from the original visual input to the hidden state H_t while keeping all other variables unchanged. This isolates the effect of visual

information on token generation by comparing the factual scenario (V intact) with the counterfactual scenario (V nullified).

Preservation of Non-Visual Context: The intervention leaves the linguistic priors L and the previous hidden state H_{t-1} unaffected. This ensures that any difference in output probabilities can be attributed solely to the removal of visual information, not to changes in language understanding or generation history.

Alignment with Structural Causal Models: In the structural equation (10), replacing V with $\mathbf{0}$ corresponds to performing a "do-operation" in the sense of Pearl's causal calculus. The resulting counterfactual hidden state $H_t^{cf} = f_\theta(\mathbf{0}, L, H_{t-1}, U)$ represents what the model's internal state would have been had the image not been present.

Interpretation of the Intervention Effect: The contrast $\Delta_t(w) = \log P_{\mathcal{F}}(w|h_t) - \log P_{\mathcal{C}}(w|h_t^{cf})$ estimates the *average causal effect* of visual information on the log-probability of generating token w. A positive $\Delta_t(w)$ indicates that the token w is visually grounded—its generation probability decreases when visual information is removed.

C.4 Limitations and Caveats

While the zeroing intervention provides a theoretically grounded approach, several practical limitations exist:

- **Non-linearity:** The effect of completely removing visual features may differ from the effect of partially altering them, especially in highly non-linear models.
- **Collider Bias:** Conditioning on the hidden state H_t (a descendant of both V and L) could potentially introduce collider bias. However, since we intervene on V before H_t is computed, this concern is mitigated.
- **Mediation vs. Confounding:** The hidden state H_t mediates the effect of V on Y_t, but may also contain confounding influences from L. Our approach separates these by comparing factual and counterfactual paths with identical L.

These considerations underscore the importance of interpreting $\Delta_t(w)$ as an *intervention effect* rather than a pure measure of visual grounding. Nevertheless, empirical results across multiple benchmarks demonstrate that this effect strongly correlates with hallucination reduction.

D Validation of the SAFE Method via Intervention Concepts

D.1 Analysis of Intervention Concepts

Definition 1 (Equations of Generation Process): Based on the method in Section 3.1, we describe the equations for the bimodal generation system. Let visual encoder ϕ map image $\mathcal{V}$ to visual features, and linguistic decoder ψ process prompts $\mathcal{P}$. The hidden state evolution follows:

$$\begin{cases} h_t = f_\theta([\phi(\mathcal{V}); \psi(\mathcal{P})], h_{t-1}) \\ h_t^{cf} = f_\theta([\mathbf{0}^{d_v}; \psi(\mathcal{P})], h_{t-1}^{cf}) \end{cases} \quad (12)$$

where $h_t \in \mathbb{R}^{d_h}$ is the factual hidden state at timestep t, h_t^{cf} is the counterfactual hidden state under intervention $do(V = \mathbf{0}^{d_v})$, and $[\cdot;\cdot]$ denotes concatenation. This aligns with the method's equations (1) and (2).

Proposition 1 (Identifiability of Counterfactual Intervention): Under certain assumptions, the intervention effect on token generation probability may be approximated through the contrast between factual and counterfactual paths, as defined in method equation (3). Specifically:

$$\Delta_t(w) = \log P_{\mathcal{F}}(w|h_t) - \log P_{\mathcal{C}}(w|h_t^{cf}) \quad (13)$$

where $P_{\mathcal{F}}$ and $P_{\mathcal{C}}$ are the generation probabilities along factual and counterfactual paths, respectively. This requires that the linguistic features $\psi(\mathcal{P})$ and generation history are consistent across paths, ensured by shared caching in the method.

Rationale: The identifiability stems from the intervention analysis framework. The intervention $do(V = \mathbf{0}^{d_v})$ severs the visual influence, allowing us to compute the intervention effect directly via the difference in log probabilities. The stability of hidden states is maintained by the recurrent nature of f_θ, and the Lipschitz continuity of f_θ (with constant L_f) ensures bounded divergence:

$$\|h_t - h_t^{cf}\| \leq L_f \|\phi(\mathcal{V})\| + \text{error terms}, \quad (14)$$

which converges under typical assumptions (e.g., $L_f < 1$). Thus, $\Delta_t(w)$ is intended to capture the visual dependency.

D.2 Consistency of Effect Measurement

Definition 2 (Effect Framework): Define the intervention effect for token w at step t as in method equation (3):

$$\Delta_t(w) = \log P_{\mathcal{F}}(w|h_t) - \log P_{\mathcal{C}}(w|h_t^{cf}). \quad (15)$$

This represents the intervention effect for token generation, where $P_{\mathcal{F}}(w|h_t)$ is the factual outcome and $P_{\mathcal{C}}(w|h_t^{cf})$ is the counterfactual outcome.

Proposition 2 (Consistency of Effect Measurement): Under ideal conditions, the measure $\Delta_t(w)$ may be consistent for the true intervention effect when the hidden states h_t and h_t^{cf} are accurately computed. Under the method's assumptions (e.g., shared linguistic features and history), $\Delta_t(w)$ may converge in probability to the true effect as the model parameters stabilize. This is supported by the law of large numbers and the continuous mapping theorem, given the smoothness of f_θ and probability functions.

D.3 Game-Theoretic Interpretation of Dynamic Penalty Mechanism

Definition 3 (Utility Function for Beam Search): Based on the dual-path dynamic penalty mechanism in method Section 3.2, define the utility for the i-th candidate sequence as the adjusted logits:

$$U_t^{(i)} = \text{logit}_{\mathcal{F}}^{(i)}(w) - \mathcal{P}_t^{(i)} \cdot \alpha_t, \quad (16)$$

where $\mathcal{P}_t^{(i)} = \lambda \cdot \exp\left(-\beta \sum_{\tau=1}^{t} s_\tau^{(i)}\right) \cdot \mathbb{I}(s_t^{(i)} < \tau_s)$ is the penalty term from method equation (5), $s_t^{(i)} = \sigma\left(\frac{1}{K}\sum_{k=t-K}^{t} \Delta_k(w^{(i)})\right)$ is the visual dependency score from method equation (4), and $\alpha_t = 1 - \frac{1}{B}\sum_{j=1}^{B} \mathbb{I}(s_t^{(j)} \geq \tau_s)$ is the competitive inhibition factor from method equation (6). This directly mirrors the method's logits adjustment.

Theorem 3 (Equilibrium in Beam Search): The beam search process with adjusted logits may approach a Bayesian Nash equilibrium where no candidate sequence can improve its utility by unilaterally changing its visual dependency. The utility function $U_t^{(i)}$ is concave in $s_t^{(i)}$ under reasonable assumptions (e.g., $\beta > 0$), suggesting the possible existence of equilibrium. Specifically:

$$\frac{\partial U_t^{(i)}}{\partial s_t^{(i)}} = \lambda\beta \exp\left(-\beta \sum_{\tau=1}^{t} s_\tau^{(i)}\right) \cdot \alpha_t \cdot \mathbb{I}(s_t^{(i)} < \tau_s) \geq 0, \quad (17)$$

indicating that increased visual dependency reduces penalty, aligning with the method's goal of promoting visually grounded tokens.

D.4 Statistical Convergence of Visual Dependency Estimation

Proposition 4 (Concentration of Visual Dependency Scores): The visual dependency score $s_t^{(i)}$ concentrates around its mean due to the averaging over a sliding window of length K. If the intervention effects $\Delta_k(w^{(i)})$ are sub-Gaussian with parameter σ^2, then for any $\epsilon > 0$:

$$P\left(\left|s_t^{(i)} - \mathbb{E}[s_t^{(i)}]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{K\epsilon^2}{2\sigma^2}\right), \quad (18)$$

where $\mathbb{E}[s_t^{(i)}]$ is the expected visual dependency. This may support the method's use of $s_t^{(i)}$ in the penalty mechanism, aiming to ensure robust detection of non-visual tokens.

D.5 Faithfulness of Structural-Aware Estimation

Definition 4 (Faithfulness Metric): The structural-aware faithfulness enhancement in method Section 3.3 aims to ensure that the generated text aligns with image structure by penalizing non-visual tokens. The fast effect test:

$$\text{is_vision}(w) = \mathbb{I}\left(\Delta(w) > \mu_\Delta + 0.5 \cdot \sigma_\Delta + 0.1\right) \quad (19)$$

identifies visually grounded tokens based on the mean μ_Δ and standard deviation σ_Δ of $\Delta_t(w)$ across tokens.

Proposition 5 (Error Bound for Hallucination Reduction): The SAFE decoding strategy aims to reduce hallucination by bounding the probability of generating non-visual tokens. Let $P_{\text{hallucination}}$ be the probability of hallucination under standard decoding, and P_{SAFE} under SAFE. Then:

$$P_{\text{SAFE}}(w) \leq P_{\text{hallucination}}(w) - \lambda \cdot \gamma^t \cdot (1 - \text{is_vision}(w)) + \text{error}, \quad (20)$$

where $\gamma^t = 1/(1 + 0.1t)$ is the decay factor. This suggests exponential decay in hallucination propensity over time, consistent with the method's objectives.

Method	Random				Popular				Adversarial			
	Acc↑	Prec↑	Recall↑	F1↑	Acc↑	Prec↑	Recall↑	F1↑	Acc↑	Prec↑	Recall	F1↑
Sample	80.61	81.03	81.46	81.25	74.56	71.06	82.86	76.51	72.10	68.44	82.00	74.61
Beam	89.31	95.97	82.73	88.86	84.33	85.51	82.66	84.06	81.90	81.54	82.46	82.00
OPERA	79.96	73.00	97.00	83.30	65.13	59.24	97.00	73.55	63.63	58.17	97.00	72.73
VCD	79.31	80.01	79.80	79.90	72.63	69.43	80.86	74.71	71.96	68.83	80.26	74.11
AGLA	82.64	84.28	81.53	82.88	76.03	73.25	82.00	77.38	74.20	70.88	82.13	76.09
SID	77.31	76.41	81.00	78.64	69.83	66.08	81.46	72.97	69.26	65.13	82.93	72.96
ICD	83.88	84.52	84.13	84.33	76.56	72.47	85.66	78.52	72.66	68.29	84.6	75.58
SAFE	87.83	97.04	78.86	86.98	84.26	88.58	78.66	83.33	81.96	83.93	79.06	81.42

Table 12: Comprehensive Performance Comparison of Various Methods on the POPE Benchmark Using InstructBLIP

Method	Art & Design↑	Business↑	Science↑	Health & Medicine↑	Human. & Social Sci.↑	Tech & Eng.↑	Overall↑
Sample	0.292	0.213	0.28	0.287	0.35	0.238	0.271
Beam	0.258	0.233	0.253	0.287	0.342	0.262	0.27
OPERA	0.325	0.273	0.227	0.313	0.383	0.243	0.287
VCD	0.217	0.227	0.24	0.267	0.325	0.248	0.252
AGLA	0.25	0.26	0.247	0.253	0.292	0.276	0.263
SID	0.283	0.3	0.253	0.213	0.342	0.276	0.276
ICD	0.217	0.213	0.2	0.3	0.325	0.324	0.267
SAFE	0.325	0.267	0.213	0.327	0.333	0.238	0.278

Table 13: Comprehensive Performance Comparison of Various Methods on the MMMU Benchmark Using InstructBLIP

E Evaluation based on InstructBLIP

As shown in Table 14, we conduct a quantitative evaluation of various hallucination detection methods using InstructBLIP, focusing on sentence-level hallucination rate (CHAIRs), instance-level hallucination rate (CHAIRi), recall, and average caption length (Len). Our proposed method, SAFE, demonstrates competitive performance across multiple metrics. Specifically, SAFE achieves a sentence-level hallucination rate of 47.0, which is lower than Sample (51.0), OPERA (55.0), VCD (57.7), AGLA (52.0), SID (50.0), and ICD (51.0), though slightly higher than Beam (42.0). For instance-level hallucination rate, SAFE matches the best-performing method Beam with a value of 12.6, significantly outperforming others such as AGLA (26.2) and SID (28.0), indicating its effectiveness in reducing fine-grained hallucinations. In terms of recall, SAFE again ties with Beam at 70.2, surpassing all other compared methods and underscoring its ability to accurately identify objects in captions. Notably, SAFE maintains a relatively short average caption length of 85.2, comparable to OPERA (85.3) and lower than methods like AGLA (108.6) and Sample (96.8), suggesting that it achieves high performance without excessive verbosity. Overall, SAFE balances hallucination suppression, recall enhancement, and caption conciseness, validating its potential as an efficient approach for hallucination detection.

Method	CHAIRs↓	CHAIRi↓	Recall↑	Len
Sample	51.0	16.0	68.3	96.8
Beam	42.0	12.6	70.2	90.6
OPERA	55.0	14.9	69.0	85.3
VCD	57.7	16.5	67.3	94.8
AGLA	52.0	26.2	64.0	108.6
SID	50.0	28.0	49.0	97.3
ICD	51.0	24.6	64.0	98.2
SAFE	47.0	12.6	70.2	85.2

Table 14: Quantitative Comparison of Different Methods for Hallucination Detection Metrics Using InstructBLIP. CHAIRs and CHAIRi represent sentence-level and instance-level hallucination rates (lower is better), Recall measures the proportion of correctly identified objects (higher is better), and Len indicates the average caption length.

As shown in Table 12, we conduct comprehensive experiments on the POPE benchmark using the InstructBLIP model, evaluating the performance

Method	CHAIRs↓	CHAIRi↓	Recall↑	Len
Sample	59.0	16.7	75.7	98.8
Beam	63.0	18.2	73.7	96.7
OPERA	49.5	15.3	66.9	93.3
VCD	50.0	15.6	67.6	93.0
SID	49.7	13.8	68.8	96.5
ICD	57.0	16.4	74.7	96.0
SAFE	60.0	13.8	72.8	96.1

Table 15: Quantitative Comparison of Different Methods for Hallucination Detection Metrics Using Shikra. CHAIRs and CHAIRi represent sentence-level and instance-level hallucination rates (lower is better), Recall measures the proportion of correctly identified objects (higher is better), and Len indicates the average caption length.

of various methods under three distinct settings: Random, Popular, and Adversarial. The evaluation metrics include Accuracy, Precision, Recall, and F1 score. The compared methods comprise Sample, Beam, OPERA, VCD, AGLA, SID, ICD, and our proposed approach, SAFE. Overall, all methods achieve the best performance in the Random setting, while their performance generally degrades in the Adversarial setting, reflecting the challenge posed by adversarial samples to model robustness. Specifically, the Beam method demonstrates high Accuracy and F1 scores in both Random and Popular settings, with an F1 score of 88.86 in the Random setting. However, the SAFE method exhibits competitive advantages across multiple metrics, particularly excelling in Precision.

In the Random setting, SAFE achieves an Accuracy of 87.83% and a Precision of 97.04%, the latter being the highest among all methods. Despite its relatively lower Recall (78.86%), which results in an F1 score (86.98%) slightly below that of Beam (88.86%), SAFE outperforms other methods in Precision. In contrast, the ICD method shows a competitive F1 score (84.33%) but lower Precision (84.52%), while the OPERA method attains a high Recall (97.00%) but a low Precision (73.00%), indicating a high false positive rate. In the Popular setting, SAFE's Accuracy (84.26%) is comparable to Beam (84.33%), and its Precision (88.58%) remains the highest, though the Recall (78.66%) limits the F1 score (83.33%) to be slightly lower than Beam (84.06%). Other methods, such as AGLA and ICD, achieve F1 scores of 77.38% and 78.52%, respectively, both inferior to SAFE. In the more challenging Adversarial setting, SAFE achieves leading Accuracy (81.96%) and Precision (83.93%), with an F1 score (81.42%) close to Beam (82.00%) and significantly higher than ICD (75.58%) and VCD (74.11%), underscoring SAFE's robustness in adversarial environments.

In summary, the SAFE method consistently leads in Precision, which helps reduce false positives, but there is room for improvement in Recall, potentially due to its conservative sample selection. Compared to existing methods, SAFE demonstrates a superior balance between Accuracy and robustness, particularly showcasing strong competitiveness in adversarial settings, thereby providing valuable insights for future research.

As shown in Table 13, we evaluate the performance of various methods on the MMMU benchmark using the InstructBLIP model across multiple subject categories, including Art & Design, Business, Science, Health & Medicine, Humanities & Social Science, Tech & Engineering, and an Overall score. The compared methods comprise Sample, Beam, OPERA, VCD, AGLA, SID, ICD, and our proposed SAFE approach. Overall, the OPERA method achieves the highest Overall score (0.287), while SAFE attains a competitive Overall score of 0.278, ranking among the top performers.

In detail, SAFE demonstrates notable strengths in specific categories: it ties for the highest score in Art & Design (0.325) with OPERA and achieves the top score in Health & Medicine (0.327). However, its performance is lower in Science (0.213) and Humanities & Social Science (0.333), where methods like OPERA and SID excel. Comparatively, SAFE's results are balanced, with no method dominating uniformly across all categories. This underscores the variability in method effectiveness depending on the domain, with SAFE showing particular promise in design- and health-related tasks while indicating areas for future improvement in science and social sciences.

F Evaluation based on Shikra

As shown in Table 15, we evaluate various methods on hallucination detection metrics using Shikra, focusing on CHAIRs (sentence-level hallucination rate), CHAIRi (instance-level hallucination rate), recall, and average caption length. Our method, SAFE, achieves competitive perfor-

Method	Random				Popular				Adversarial			
	Acc↑	Prec↑	Recall↑	F1↑	Acc↑	Prec↑	Recall↑	F1↑	Acc↑	Prec↑	Recall	F1↑
Sample	85.46	86.16	85.53	85.84	83.73	82.56	85.53	84.02	80.36	77.52	85.53	81.33
Beam	87.11	90.73	83.53	86.92	84.10	84.49	83.53	84.00	81.70	80.57	83.53	82.02
OPERA	81.95	83.41	81.13	82.25	81.20	81.24	81.13	81.18	76.60	74.38	81.13	77.61
VCD	81.99	83.65	80.86	82.23	80.90	80.92	80.86	80.89	76.60	74.47	80.93	77.57
SID	80.13	80.01	81.93	80.96	77.43	76.02	80.13	78.02	73.83	70.62	81.60	75.71
ICD	80.92	80.30	83.46	81.85	79.53	77.41	83.40	80.29	76.26	72.54	84.53	78.07
SAFE	85.15	85.31	86.00	85.65	83.40	81.66	86.13	83.84	79.83	76.62	85.86	80.98

Table 16: Comprehensive Performance Comparison of Various Methods on the POPE Benchmark Using Shikra

Method	Art & Design↑	Business↑	Science↑	Health & Medicine↑	Human. & Social Sci.↑	Tech & Eng.↑	Overall↑
Sample	0.242	0.253	0.207	0.367	0.317	0.295	0.281
Beam	0.15	0.28	0.3	0.287	0.242	0.31	0.269
OPERA	0.292	0.233	0.187	0.287	0.25	0.257	0.25
VCD	0.267	0.247	0.2	0.28	0.217	0.3	0.256
SID	0.258	0.253	0.207	0.26	0.258	0.243	0.246
ICD	0.258	0.193	0.333	0.233	0.283	0.229	0.252
SAFE	0.275	0.24	0.247	0.347	0.233	0.338	0.29

Table 17: Comprehensive Performance Comparison of Various Methods on the MMMU Benchmark Using Shikra

mance, notably tying for the lowest CHAIRi score of 13.8 with SID, indicating strong effectiveness in reducing instance-level hallucinations. However, SAFE shows a higher CHAIRs score of 60.0 compared to top performers like OPERA (49.5) and SID (49.7), suggesting room for improvement in sentence-level hallucination control. In terms of recall, SAFE attains 72.8, which is lower than Sample (75.7) but comparable to Beam (73.7) and ICD (74.7), while maintaining a similar caption length (96.1) to other methods. Overall, SAFE demonstrates a balanced trade-off, excelling in instance-level hallucination reduction while remaining competitive in recall and caption coherence.

As shown in Table 16, we evaluate various methods on the POPE Benchmark using Shikra, assessing accuracy, precision, recall, and F1 score across Random, Popular, and Adversarial splits. Our method, SAFE, delivers competitive performance, achieving F1 scores of 85.65, 83.84, and 80.98 in the respective splits. Notably, SAFE consistently attains high recall values (86.00, 86.13, 85.86) across all conditions, surpassing or matching leading baselines such as Beam and Sample in this metric, which underscores its effectiveness in robust object identification. However, in the Adversarial split, SAFE's precision (76.62) lags behind Beam (80.57), contributing to a marginally lower F1 score. Overall, SAFE demonstrates a balanced trade-off, excelling in recall-driven scenarios while maintaining competitive accuracy and F1, though further refinement could enhance precision in adversarial settings.

As shown in Table 17, we evaluate various methods on the MMMU Benchmark using Shikra, assessing performance across diverse domains including Art & Design, Business, Science, Health & Medicine, Humanities & Social Sciences, and Tech & Engineering. Our method, SAFE, achieves the highest overall score of 0.29, surpassing competitive baselines such as Sample (0.281) and Beam (0.269). Notably, SAFE excels in the Tech & Engineering domain with a score of 0.338, the best among all methods, and demonstrates strong performance in Health & Medicine (0.347), closely trailing Sample (0.367). However, SAFE shows relatively lower results in Humanities & Social Sciences (0.233) and Science (0.247), where it underperforms compared to methods like ICD (0.333 in Science). Overall, SAFE maintains a balanced and competitive profile, particularly in technical domains, while highlighting areas for future refinement in social and scientific reasoning.

Method	Total Time	Average Time	Standard Deviation	Shortest Time	Longest Time
Sample	674.94	11.24	7.01	0.98	26.21
Beam	2032.92	33.88	20.64	2.80	77.23
OPERA	2484.19	41.40	24.73	3.32	89.99
VCD	1465.08	24.41	14.28	0.67	54.08
SID	2328.94	38.81	23.89	3.22	87.10
ICD	2288.95	38.14	24.66	3.23	87.16
AGLA	2215.34	36.92	24.24	2.94	79.76
SAFE	2405.20	77.58	36.60	3.18	131.84

Table 18: Time complexity analysis. All times are in seconds.

G Time Complexity Analysis

In our time complexity analysis conducted on the llava-bench-in-the-wild benchmark, we assess the computational efficiency of various hallucination suppression methods under identical decoding configurations, including a temperature of 1.0, beam size of 3, and maximum generation length of 256. As detailed in Table 18, our method, SAFE, incurs the highest average inference time of 77.58 seconds, with a total time of 2405.20 seconds and the longest single inference duration of 131.84 seconds, reflecting the computational overhead introduced by its additional intervention analysis path for token-level hallucination detection and suppression. In comparison, baseline methods such as Sample (average time: 11.24 seconds) and VCD (average time: 24.41 seconds) demonstrate significantly lower latency, while other approaches like OPERA (41.40 seconds) and SID (38.81 seconds) also show moderate efficiency. The standard deviation for SAFE (36.60 seconds) is notably higher, indicating greater variability in processing times, which may stem from the dynamic nature of its token-level operations. Although this increased cost aligns with SAFE's enhanced performance in reducing hallucinations as evidenced in prior experiments, it underscores a key trade-off between accuracy and computational resources, suggesting avenues for future optimization to balance effectiveness with practical deployment constraints.

H Ablation Study

We analyze the ablation study results presented in Tables 19, 20, and 21. First, we note that all thresholds (including effectiveness threshold, confidence threshold, and t-statistic threshold) are determined using an equal-interval tri-section method based on the observed minimum and maximum values from experiments, with three equally-spaced threshold points evaluated for each.

Effectiveness Threshold	**qAcc↑**	**fAcc↑**
0.1	13.84	12.42
0.3	12.30	13.29
0.5	10.76	14.73

Table 19: Ablation Study: Effectiveness Threshold

Confidence Threshold	**qAcc↑**	**fAcc↑**
0.5	13.84	12.42
0.6	11.86	12.71
0.7	10.54	13.58

Table 20: Ablation Study: Confidence Threshold

For the effectiveness threshold (Table 19), as the threshold increases from 0.1 to 0.5, qAcc decreases from 13.84 to 10.76, while fAcc increases from 12.42 to 14.73. This indicates that higher effectiveness thresholds favor fAcc at the expense of qAcc, reflecting a trade-off between these two metrics.

For the confidence threshold (Table 20), when the threshold rises from 0.5 to 0.7, qAcc declines from 13.84 to 10.54, whereas fAcc improves from 12.42 to 13.58. Similar to the effectiveness threshold, higher confidence thresholds benefit fAcc while reducing qAcc, further confirming the impact of threshold selection on performance balance.

For the t-statistic threshold (Table 21), as the

t-Statistic Threshold	**qAcc↑**	**fAcc↑**
100	13.84	12.42
200	10.54	10.40
300	10.10	12.13

Table 21: Ablation Study: t-Statistic Threshold

threshold increases from 100 to 300, qAcc consistently decreases from 13.84 to 10.10. In contrast, fAcc first drops to its lowest value of 10.40 at a threshold of 200, then recovers to 12.13 at 300. This suggests that the t-statistic threshold's effect on fAcc is non-monotonic, requiring careful selection based on task characteristics.

Overall, the experimental results across Tables 19, 20, and 21 reveal consistent trends: increasing thresholds generally lead to decreased qAcc, while fAcc exhibits either improvement or a more complex pattern of initial decline followed by recovery. These findings provide valuable insights for threshold optimization, demonstrating that the equal-interval tri-section method effectively reveals how threshold variations influence model performance.

I Ethical Considerations

The development and deployment of hallucination mitigation techniques like SAFE raise several ethical considerations that warrant careful discussion:

Positive Impacts on AI Safety and Reliability: By reducing hallucinations in large vision-language models, SAFE contributes to building more trustworthy and reliable AI systems. This is particularly important for real-world applications where accurate visual understanding is critical, such as medical image analysis, autonomous driving, and assistive technologies for visually impaired individuals. Reducing factual inconsistencies helps prevent harmful decisions based on erroneous model outputs.

Computational and Environmental Costs: As noted in the time complexity analysis (Section 18), SAFE introduces significant computational overhead due to its dual-path architecture. This increased energy consumption has environmental implications, particularly when deployed at scale. Researchers and practitioners should consider this trade-off between accuracy improvements and sustainability, potentially exploring optimization strategies or selective application in critical scenarios.

Potential for Misuse: While SAFE aims to enhance factual accuracy, any technology that makes AI outputs more credible could potentially be misused to generate more convincing disinformation or deepfakes. The same mechanisms that suppress hallucinations in legitimate applications could theoretically be adapted to create more persuasive synthetic media. We emphasize that SAFE should be deployed responsibly with appropriate safeguards against malicious use.

Transparency and Interpretability: SAFE's intervention-inspired approach provides some level of interpretability through token-level visual dependency scores. However, the complexity of the underlying mechanisms may limit comprehensibility for non-expert users. Future work should focus on improving explainability and developing user-friendly interfaces that help stakeholders understand when and why the model makes specific decisions.

Bias and Fairness Considerations: Hallucination mitigation techniques may interact with existing biases in training data or model architectures. If visual dependency estimation systematically favors certain types of content or representations, it could inadvertently amplify disparities. While our current evaluation focuses on accuracy metrics, future research should examine potential differential impacts across demographic groups, cultural contexts, and content categories.

Deployment Considerations: The practical implementation of SAFE requires careful consideration of application contexts. In high-stakes domains like healthcare or legal analysis, even reduced hallucination rates may still pose risks. We recommend thorough testing in domain-specific settings, clear communication of system limitations to end-users, and maintaining human oversight for critical decisions.

In conclusion, while SAFE represents a technical advancement in hallucination mitigation, its ethical implications extend beyond algorithmic performance. We encourage the research community to engage in ongoing dialogue about responsible development, deployment, and governance of such technologies, balancing innovation with consideration of broader societal impacts.

J Evaluation Protocol

We conduct experiments using LLaVA-1.5, InstructBLIP, and Shikra models, with weights obtained directly from the Hugging Face community without additional training or quantization. Model weights are loaded via the Transformers library (version 4.57.1), and inference is performed with float16 precision using serial (non-batched) processing. We maintain consistent decoding configurations including beam search, temperature, top-p,

and top-k parameters.

For all benchmarks, we adopt their original evaluation protocols. For HallusionBench, MMHal-Bench, POPE, CHAIR, and MMMU, we use consistent prompt templates, the GPT-OSS-20B evaluation model, and evaluation functions, all sourced from HallusionBench.

Algorithm 1 SAFE: Structural-Aware Faithfulness Enhancement for Hallucination Mitigation

Require: Vision-language model $\mathcal{M}$, image $\mathcal{V}$, prompt $\mathcal{P}$, decoding mode $m \in \{\text{beam}, \text{sample}\}$

Ensure: Generated text sequence $W = w_1, w_2, \ldots, w_T$

Initialize factual hidden state h_0 and counterfactual hidden state h_0^{cf}
Initialize generated sequence $W \leftarrow \emptyset$
for $t = 1$ to T **do**
  Compute factual logits $\mathbf{L}_f$ using h_{t-1} and visual input $\phi(\mathcal{V})$
  Compute counterfactual logits $\mathbf{L}_c$ using h_{t-1}^{cf} and masked visual input $\mathbf{0}^{d_v}$
  Compute intervention effect $\Delta_t = \mathbf{L}_f - \mathbf{L}_c$
  if $m = \text{beam}$ **then**
    Apply dual-path dynamic penalty mechanism (Section 3.2):
    Compute visual dependency scores $s_t^{(i)}$ for each beam candidate
    Compute penalty terms $\mathcal{P}_t^{(i)}$ and competitive inhibition factor α_t
    Adjust logits: $\text{logit}_{\text{adj}}^{(i)} = \text{logit}_{\mathcal{F}}^{(i)} - \mathcal{P}_t^{(i)} \cdot \alpha_t$
  else
    (standard sampling) Apply structural-aware faithfulness enhancement (Section 3.3): Compute fast effect test for each token w: $\text{is_vision}(w) = \mathbb{I}(\Delta_t(w) > \mu_\Delta + 0.5 \cdot \sigma_\Delta + 0.1)$ Adjust logits for each token w: $\mathbf{L}_{\text{adjusted}}(w) = \mathbf{L}_f(w) - \lambda \cdot \gamma^t \cdot (1 - \text{is_vision}(w))$
  end if
  Sample next token w_t from adjusted logits distribution
  Append w_t to W
  Update hidden states h_t and h_t^{cf} for next step
end for
return W