What Do Hallucinations Reveal About Multimodal Reasoning? Diagnosing Visual Grounding Failures via Contrastive Decoding Probes

Zhipeng Zhao[1], Wenxu Wang[1], Peishun Liu[1], Ruichun Tang[1,*]

[1]Ocean University of China

{zhaozhipeng}@stu.ouc.edu.cn

{wangwenxu, liups, tangruichun}@ouc.edu.cn

[*]Corresponding author

Abstract

When strong multimodal models are widely available, progress requires new scientific methodologies beyond benchmark scores—using models as instruments for understanding behavior. We address this by asking: can we use large vision-language models (LVLMs) as experimental instruments for studying their own failure dynamics? Focusing on visual hallucination, we introduce SAFE, a training-free decoding framework that contrasts visually-grounded and vision-ablated generation paths to produce a token-level contrastive grounding score that identifies when the model favors linguistic priors over visual evidence. This signal serves dual roles: as a practical proxy for detecting visually-ungrounded tokens, and as the basis for decoding-time penalties. Our analysis yields three empirical observations: visual dependency decays over generation, hallucinations co-occur in temporal clusters, and early intervention reduces clustering without substantially degrading fluency. On MMHal-Bench, SAFE substantially outperforms all compared baselines; results elsewhere are more mixed. We argue that designing contrastive probes exemplifies a broader mission: using models as instruments for scientific understanding. Code: https://github.com/zhaozhipeng1997/SAFE_public.

1 Introduction

When strong general-purpose models are increasingly available, a foundational question arises: what should the missions of NLP research be? One answer, which we pursue here, is that progress demands new scientific methodologies—using models as experimental instruments to probe *how* and *why* they behave, rather than treating them only as systems to optimize against static leaderboards.

We apply this perspective to visual-language hallucination in LVLMs (Liu et al., 2023; Bai et al., 2025; Wu et al., 2024). As illustrated in Figure 1, even strong models routinely generate text contradicting visual input, eroding trustworthiness.

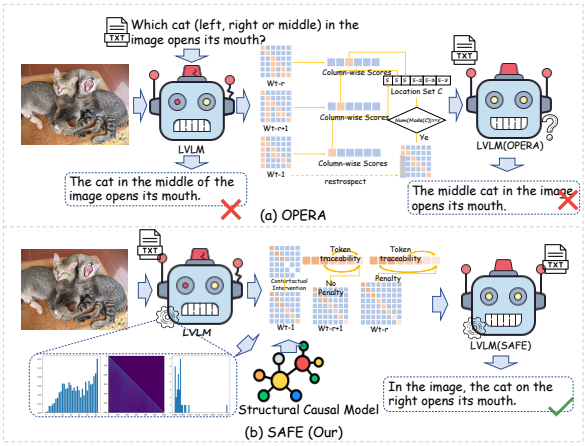


Figure 1: Comparison of hallucination mitigation paradigms and SAFE.

Current hallucination research is dominated by two paradigms: training-time interventions (Chen et al., 2025; Zhang et al., 2025a) and post-hoc decoding heuristics (Liu et al., 2025; Huang et al., 2024). Both improve benchmarks but treat hallucination as a defect to suppress rather than a phenomenon to study. We argue that understanding *how* models negotiate between visual evidence and linguistic priors requires a diagnostic instrument—not another mitigation technique—that quantifies token-level visual grounding and reveals temporal failure dynamics.

We introduce contrastive probing: by running generation paths with and without visual input and computing the token-level log-probability gap, we obtain a **contrastive grounding score** as a real-time proxy for visual dependency. We emphasize this is a *diagnostic heuristic*, not formal causal identification. We call the framework **SAFE** (**S**tructural-**A**ware **F**aithfulness **E**nhancement), operating along two dimensions:

- **As a diagnostic probe:** Token-level visual

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 2889–2912
October 24-29, 2026 ©2026 Association for Computational Linguistics

dependency scores identify which tokens lack grounding and when language priors dominate.

- **As a mitigation method:** SAFE penalizes low-dependency tokens via dual-path contrast with adaptive penalty decay and competitive inhibition.

Across five benchmarks and three 7B architectures, visual dependency decays over decoding, hallucinations cluster temporally, and early intervention reduces clustering without substantially degrading fluency. The method is particularly effective on open-ended hallucination benchmarks (MMHal-Bench), with mixed results where external knowledge or visual ambiguity dominates. This work illustrates a research direction we argue is increasingly central: designing contrastive probes that transform models into instruments for scientific understanding.

2 Related Work

LVLMs such as LLaVA-1.5 (Liu et al., 2023), InstructBLIP (Dai et al., 2023), and Shikra (Chen et al., 2023) integrate LLMs (Guo et al., 2025; OpenAI et al., 2024) with visual encoders, yet all exhibit visual-language hallucinations (Zhang et al., 2025b; Tonmoy et al., 2024). This pervasiveness suggests hallucination is rooted in how models balance visual evidence against strong language priors.

2.1 Hallucination Mitigation: Three Analytical Paradigms

Training-based alignment. Methods fine-tune models on curated data to strengthen visual-semantic correspondence (Chen et al., 2025; Zhang et al., 2025a) or modify architectures for visual feature integration (Xie et al., 2024; Shang et al., 2025). While effective given high-quality data, they obscure *why* hallucination rates decrease and generalize poorly to unseen visual concepts.

Heuristic decoding interventions. OPERA (Huang et al., 2024) penalizes over-reliance on summary tokens. Contrastive decoding (O'Brien and Lewis, 2023) has spawned multimodal variants: VCD (Leng et al., 2024), ICD (Wang et al., 2024), ConVis (Park et al., 2025), LCD (Manevich and Tsarfaty, 2024), CATCH (Kan et al., 2024), IFCD (Wang et al., 2025a), SDCD (Xia et al., 2026), ASCD (Wang et al., 2026). SIRA (Qin et al., 2026) constructs internal counterfactuals via attention masking. DCD (Wang et al., 2025b) and INTER (Dong et al., 2025) correct decoding online. Dropout Decoding (Fang et al., 2025b) and ECD (Fieback et al., 2025) use uncertainty masking. Calibration by Fang et al. (2025a) targets visual-language balance; ClearSight (Yin et al., 2025) amplifies visual signals. Grounding-score methods include VGS-Decoding (Kolli et al., 2026) and IECD^2 (Bangde and Roy, 2026). CLIP-guided (Deng et al., 2024), GLSim (Park and Li, 2025), and HACL (Jiang et al., 2024) use auxiliary vision models or contrastive learning. Domain extensions: Med-VCD (Mahdavi et al., 2026), 3D-VCD (Ogunleye et al., 2026). SAFE's contribution is the dual-forward-pass log-probability gap with parameterized penalty and diagnostic analysis.

Token-level visual diagnostics. VISTA (Li et al., 2025) analyzes visual information decay with token-level steering. TruthPrInt (Duan et al., 2025) uses latent truthful signals as per-token indicators. Cao et al. (2026) and Nguyen et al. (2026) examine attention structure and fine-grained token grounding for hallucination detection. These share SAFE's premise; SAFE's distinction is the log-probability gap between dual passes paired with penalty-based control.

Diagnosis-informed decoding. Prior paradigms do not systematically probe *when* visual evidence is abandoned or *how* ungrounded tokens affect generation. Our contrastive probing quantifies token-level visual dependency. Geigle et al. (2024) show stronger grounding does not always reduce hallucination, contextualizing SAFE's mixed results. The key distinction is that the same signal drives analysis and mitigation within a unified framework.

2.2 Empirical Diagnosis via Contrastive Probing

We connect to controlled manipulation for probing neural models: counterfactual analysis has examined linguistic representations, and perturbation-based methods have studied vision-language relationships—mostly for post-hoc explanation, not real-time diagnosis. We operationalize contrastive probing—comparing behavior with and without a targeted input—as a real-time signal embedded in decoding.

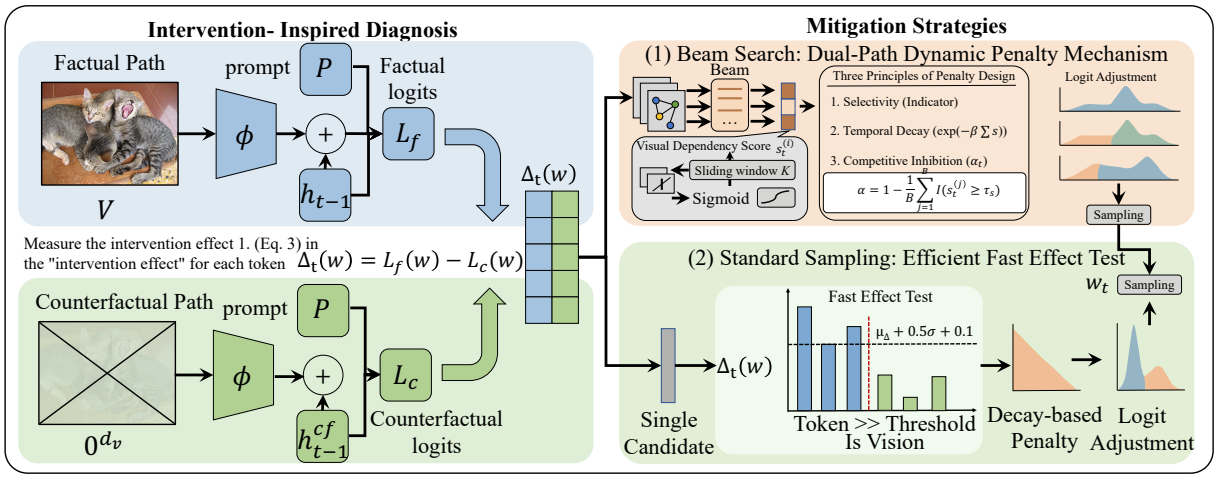


Figure 2: SAFE dual-path decoding: visually-grounded vs vision-ablated paths with progressive grounding penalty.

3 Method

3.1 The Central Question: How to Measure a Token's Visual Grounding?

To what extent does a token depend on visual information versus linguistic priors? Answering this requires isolating visual contributions—a contrastive diagnostic probe. Let the LVLM $\mathcal{M}$ be decomposed into a visual encoder $\phi : \mathcal{V} \to \mathbb{R}^{d_v}$ and language decoder $\psi : \mathcal{P} \to \mathbb{R}^{d_l}$. At time step t, the hidden state $h_t \in \mathbb{R}^{d_h}$ evolves as:

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

where ϕ maps image $\mathcal{V}$ to visual features, ψ processes textual prompts $\mathcal{P}$, $[\cdot;\cdot]$ denotes concatenation, and f_θ is a parameterized Transformer layer.

Equation (1) cannot separate visual from linguistic contributions. To isolate the visual contribution, we create a diagnostic contrast by zeroing visual features while preserving all other components (see Appendix E for conceptual motivation and discussion of alternative ablation strategies). We use *do*-notation descriptively, by analogy to causal formalisms, without implying satisfaction of formal identification conditions. This yields a vision-ablated state:

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

where $\mathbf{0}^{d_v}$ replaces visual features with a zero vector of the same dimension d_v as $\phi(\mathcal{V})$. The two paths differ *only* in visual input, isolating the visual contribution to token probabilities. We choose zero ablation for determinism, maximal contrast, and preserved linguistic context. Compared to learned null embeddings (MMHalBench 3.48 vs. 3.55), shuffled tokens (3.22), and Gaussian noise (3.15), zero ablation performs comparably to learned alternatives while being simpler (full details in Appendix E). The **diagnostic contrast score** is the log-probability difference:

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

where $P_{\mathcal{F}}$ and $P_{\mathcal{C}}$ are visually-grounded and vision-ablated probabilities. $\Delta_t(w) > 0$ means visual information increases the token's probability (visually grounded); $\Delta_t(w) \approx 0$ signals linguistic-prior dominance. This scalar, computed at every step for every candidate, is SAFE's core diagnostic signal. We emphasize that $\Delta_t(w)$ is a diagnostic statistic whose validity rests on its empirical utility for detecting and suppressing hallucinations, not on satisfaction of formal causal identification assumptions.

3.2 From Diagnosis to Mitigation: The Dual-Path Penalty Mechanism

$\Delta_t(w)$ provides a per-token grounding estimate. We translate this into mitigation via three principles: sliding-window aggregation for robustness; time-sensitive penalties—strong early, decaying as context accumulates; and global relaxation when most candidates are well-grounded. The penalty is:

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

where $\sigma(\cdot)$ normalizes to $(0,1)$ and $\min(K,t)$ handles boundaries. We design three penalty properties: (i) **Selectivity**—only tokens below τ_s penalized; (ii) **Temporal decay**—$\exp(-\beta \sum s)$ weakens penalty as visual evidence accumulates; (iii)

Competitive inhibition—α_t relaxes when candidates are well-grounded.

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

where λ, β, τ_s are penalty strength, decay rate, and threshold. The indicator, exponential, and α_t (Equation 6) implement the three properties. The final logit adjustment with competitive inhibition is:

$$\log P_{\text{adj}}^{(i)}(w) = \log P_{\mathcal{F}}^{(i)}(w) - \mathcal{P}_t^{(i)} \cdot \alpha_t$$
$$\alpha_t = 1 - \frac{1}{B} \sum_{j=1}^{B} \mathbb{I}(s_t^{(j)} \geq \tau_s) \quad (6)$$

where α_t decreases as more candidates meet the threshold, relaxing the global penalty when visual dependency is already strong. For reproducibility, our default hyperparameters are: window length K=5, penalty strength λ=2.0, decay rate β=0.1, threshold τ_s=0.3, beam size B=3, and temperature 1.0. Sensitivity to these choices is reported in Appendix J.

3.3 Two Variants of the Same Diagnostic Principle

The full mechanism targets beam search. For sampling, a lightweight variant (Section 3.4) preserves the dual-path contrast with a fast effect test. Beam uses sliding-window aggregation, exponential decay, and competitive inhibition; sampling uses a single global threshold with simple decay. We treat them as related heuristics within a common framework.

3.4 Efficient Approximation for Standard Sampling

For standard sampling, we present a lightweight variant preserving the dual-path contrast while replacing the penalty machinery with a fast effect test (see Appendix E):

$$P(Y \mid X \text{ ablated}) = \sum_{Z} P(Y \mid X, Z) P(Z) \quad (7)$$

illustrating the contrastive motivation with shared language context Z, though exact marginalization is intractable. We approximate by computing $\Delta = \mathbf{L}_f - \mathbf{L}_c$ per step with shared history. The diagnostic test uses a threshold motivated by Cohen's d:

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

where $0.5 \cdot \sigma_\Delta$ corresponds to Cohen's d effect size (Cohen, 2013); 0.1 prevents false positives when $\sigma_\Delta \approx 0$ (Appendix D). Non-visual tokens are penalized analogously:

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

with $\gamma^t = 1/(1 + 0.1t)$ mirroring temporal decay. This preserves the core principle—measure visual dependency, then intervene—with reduced overhead.

3.5 The Unified SAFE Decoding Algorithm

Algorithm 1 unifies diagnosis and mitigation: at each step, the diagnostic contrast Δ_t is computed first, then the penalty mechanism (beam search) or fast effect test (sampling) is applied. The algorithm operates as follows:

4 Experiment

We ask two questions: do SAFE's diagnostic signals improve benchmark performance, and what patterns do internal signals reveal? We validate on benchmarks (Section 4.2), then analyze token-level dependency, temporal clustering, and fluency (Sections 4.2.1–4.2.3). Default hyperparameters (K=5, λ=2.0, β=0.1, τ_s=0.3, B=3, temperature 1.0) are selected on a validation split and used uniformly across all models and benchmarks; sensitivity analysis is in the appendix. SAFE incurs $\sim$2$\times$ the inference cost of comparable methods (Table 16, Limitations).

4.1 Benchmarks and Experimental Setup

We evaluate on five benchmarks: HallusionBench (Guan et al., 2024) (language hallucinations and visual illusions), MMHalBench (Sun et al., 2024) (open-ended hallucination), CHAIR (Rohrbach et al., 2018) (object hallucination at sentence/instance levels), POPE (Li et al., 2023) (object existence), and MMMU (Yue et al., 2024) (multimodal reasoning). Beam search for MMHalBench, CHAIR, POPE; sampling for HallusionBench and MMMU. Baselines: Sample, Beam, OPERA (Huang et al., 2024), VCD (Leng et al., 2024), AGLA (An et al., 2025), SID (Huo et al., 2025), ICD (Wang et al., 2024) on LLaVA-1.5 (Liu et al., 2023), InstructBLIP (Dai et al., 2023), Shikra (Chen et al., 2023) (all 7B).

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 1: Performance on HallusionBench using LLaVA-1.5. qAcc: Question Pair Accuracy; fAcc: Figure Accuracy; LH: Language Hallucination; VI: Visual Illusion; C/I/W: Correct/Inconsistent/Wrong.

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 2: POPE results on LLaVA-1.5

4.2 Results

We test whether SAFE's detected patterns correspond to standard benchmark penalties. All experiments: LLaVA-1.5, InstructBLIP, Shikra (7B), temperature 1.0, beam size 3. **HallusionBench.** HallusionBench includes visual illusions where visual reliance can produce errors, making "visual dependency" ambiguous. SAFE achieves competitive qAcc (14.50) and fAcc (17.34), but VCD has higher qAcc (14.72) and OPERA comparable fAcc (17.05). This is expected: dependency diagnosis works best with reliable visual evidence.

MMHalBench. SAFE achieves overall score 3.55—more than double the nearest baseline (OPERA, 1.69)—with hallucination ratio 0.44 versus 0.74–0.83 (Table 3). Gains are largest in Relation (4.25) and Environment (4.42), categories penalizing visually inconsistent claims about spatial and contextual relationships—precisely SAFE's target. Unlike CHAIR (object-level string matching) or POPE (binary questions), MMHalBench uses GPT-based evaluation of free-form answers, making it sensitive to visual-semantic consistency across multi-token descriptions. This gain warrants caution: MMHalBench relies on a single GPT-based judge over 96 samples, and SAFE's conservative generation may be scored favorably in this rubric. We treat this as strong but benchmark-specific evidence.

CHAIR. SAFE achieves CHAIRi 14.8 (competitive with OPERA's 14.5) and recall 74.6 (Table 5), but OPERA is better on CHAIRs (50.5 vs. 53.0). On Shikra, SAFE ties best CHAIRi (13.8) but worst CHAIRs (60.0) (Appendix Tables 12–13). Instance-level strength is expected—token diagnosis detects unsupported objects—while sentence-level gap reflects discourse planning beyond per-token scope. SAFE generates shorter captions (e.g., 95.3 vs. 101.8 for LLaVA-1.5); examined in Section 4.2.1.

POPE. SAFE achieves highest precision (95.94 Random, 89.98 Popular) with stable Recall (82.06), confirming the penalty targets visually unsupported tokens. Adversarial results (83.00/82.06) mirror this balanced pattern. See Figure 3.

MMMU. MMMU reveals the *boundary* of visual dependency diagnosis: SAFE's overall score (0.341 vs. Beam 0.338) is marginally higher but inconsistent. Humanities & Social Sciences (0.517) benefits from visual interpretation, while Science (0.233 vs. Sample 0.320) shows the method cannot help when answers require non-visual knowledge. MMMU illustrates a clear domain boundary. See Figure 4 and appendix.

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 3: Performance on MMHalBench using LLaVA-1.5. HR: Hallucination Ratio; Attri: Attribute; Adver: Adversarial; Comp: Comparison; Env: Environment.

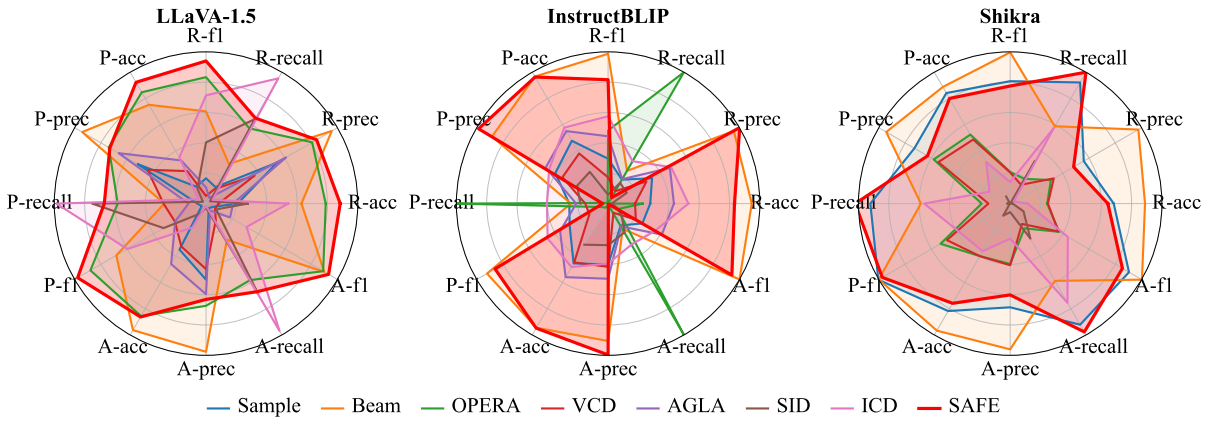


Figure 3: POPE results across models. R/P/A: Random/Popular/Adversarial.

4.2.1 Token-Level Visual Dependency Analysis

Table 6 reports visually associated tokens—tokens where Δ_t indicates visual grounding—with and without SAFE. SAFE increases the proportion of visually grounded tokens (HB: 5.43%→5.53%; MB: 1.13%→1.41%) while total tokens *decreases* (HB: 31,814→31,778; MB: 3,715→3,546). This suggests SAFE selectively suppresses low-dependency tokens, yielding shorter, more visually-attentive outputs.

To verify the "visual dependency decay" claim, we collected per-step Δ_t trajectories via dual-path inference on COCO images. Figure 5 shows aggregate trajectories: LLaVA Δ_t declines from 0.39 (first half) to 0.07 (second half); Shikra from 1.13 to 0.41, providing direct evidence that visual grounding weakens as language priors accumulate.[1]

Length-controlled analysis. To address the concern that SAFE's gains may reflect output shortening, we compute length-normalized $\text{CHAIRs}_{\text{norm}}$ = CHAIRs/mean caption length and hallucination rate per 100 tokens. On LLaVA-1.5, SAFE achieves $\text{CHAIRs}_{\text{norm}}$=0.60% (Beam 0.62%, Sample 0.65%), 16.0 hallucinated tokens/100 (Beam 16.2, Sample 17.8). Improvement direction persists after normalization, though margins are modest. On Shikra, OPERA (0.59%) and VCD (0.60%) outperform SAFE (0.74%) on $\text{CHAIRs}_{\text{norm}}$, consistent with weaker sentence-level performance on this architecture. Full results in Appendix A.

4.2.2 Validating the Diagnostic Signal

To verify that Δ_t captures grounding-specific information, we compare it against token-level entropy H and confidence $\max p(w)$ for predicting CHAIR hallucination labels on 50 COCO images (same beam-search sequences). Δ_t achieves AUROC 0.588, outperforming entropy (0.380) and confidence (0.401), both of which fall *below* random (0.50). A calibration analysis confirms reliability: hallucination rate decreases monotonically from 1.60% (lowest Δ_t decile) to 0.43% (ninth decile, slope -0.019, Figure 6). The modest absolute AUROC reflects CHAIR's annotation sparsity (∼1.2%

[1] InstructBLIP's Q-Former produces $\Delta_t \approx 0$ because learned queries dominate cross-modal fusion; trajectories are informative for LLaVA and Shikra.

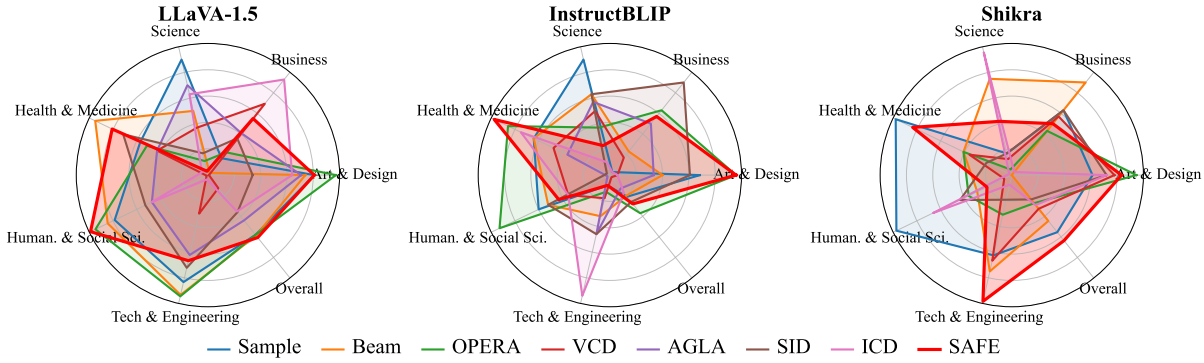


Figure 4: MMMU results across models.

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: MMMU results on 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 on CHAIR using LLaVA-1.5. CHAIRs/CHAIRi: sentence/instance-level hallucination (↓); Recall: correct objects (↑); Len: average caption length.

	Setting	TNS	TNT$_\downarrow$	NVAT$_\uparrow$	ARVAT$_\uparrow$
HB	w/o SAFE	951	31814	1727	5.43%
	w/ SAFE	951	31778	**1757**	**5.53%**
MB	w/o SAFE	96	3715	42	1.13%
	w/ SAFE	96	3546	**50**	**1.41%**

Table 6: 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 (MMHalBench). Setting indicates with (w/) or without (w/o) SAFE.

object-level labels) rather than poor signal quality. On 100 COCO images (9,397 pairs), grounded Δ_t=0.211 vs. hallucinated 0.185 (AUROC 0.574). Sentence-level aggregation fails (AUROC 0.464, Appendix B).

4.2.3 Does Suppressing Hallucinations Compromise Fluency?

Table 7 reports perplexity under Qwen2-7B and GPT-2-medium. SAFE achieves lower perplexity than Sample and Beam on MMHalBench (PPL1: 10.23 vs. 11.75/10.50), competitive on HallusionBench. This is tentative evidence that removing ungrounded tokens improves coherence, though external LM perplexity is a rough proxy for multimodal outputs. Stronger claims require human evaluation.

4.2.4 The Chain Structure of Hallucinations

Table 8 compares SAFE against a single-window baseline on propagative hallucination probability. SAFE reduces temporal clustering by 38.0% (HallusionBench) and 18.2% (MMHalBench). We caution P_{prop} measures co-occurrence, not causal propagation. SAFE's sliding window detects low-dependency clusters, and the cumulative penalty provides escalating pressure. See Appendix C.

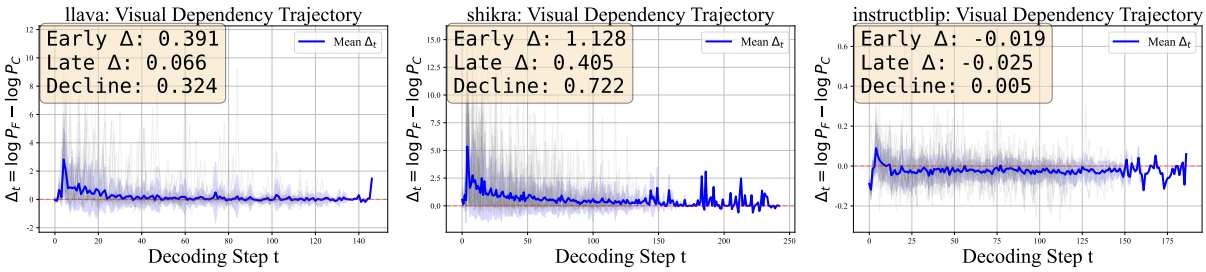


Figure 5: Δ_t trajectories on 30 COCO images. Gray: individual samples; blue: mean ± 1 std; red dashed: $\Delta_t = 0$. LLaVA Δ_t declines from 0.39 (early) to 0.07 (late); Shikra from 1.13 to 0.41.

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 7: 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 8: Evaluation of Propagative Hallucination. P_{prop} denotes propagative hallucination probability (the likelihood that an error at step t is followed by errors in subsequent steps), while N_{primary} represents the count of antecedent errors.

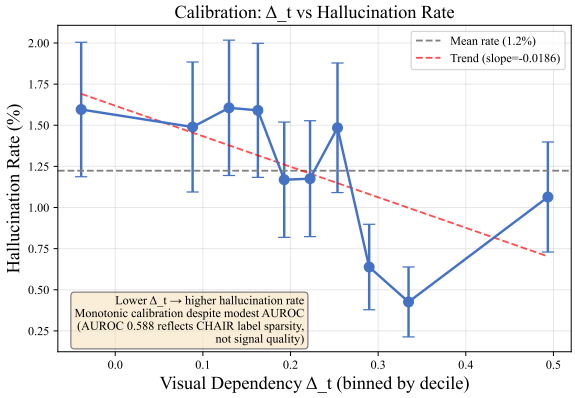


Figure 6: Calibration of Δ_t against CHAIR hallucination labels. Tokens binned by Δ_t decile; shaded band shows ± 1 standard error. Hallucination rate declines monotonically as Δ_t increases (trend slope -0.019), confirming the signal is well-calibrated despite modest absolute AUROC.

5 Conclusion

This paper asked: *what are the missions of NLP research when strong models are widely available?* Our answer, via SAFE, is that designing instruments to probe *how* models work—not just measuring *whether* they work—is a central mission. Token-level contrastive probing reveals that visual grounding decays over generation, hallucinations co-occur temporally, and early intervention reduces clustering without substantially degrading fluency. The method is effective on open-ended hallucination (MMHalBench), with mixed results elsewhere—informing *when* token-level diagnosis helps. SAFE demonstrates that using models as experimental instruments yields insights complementing architectural innovation and benchmark optimization, representing a productive direction for NLP.

Limitations

SAFE incurs $\sim 2\times$ the inference cost of comparable methods (Table 16); this could be reduced through shared KV-caching between paths. $\Delta_t(w)$ captures association between visual input and token probability rather than a formal causal mechanism; its interpretation as "visual dependency" is validated by empirical hallucination reduction, not causal proofs. The visual ablation (zeroing) assumes clean visual-linguistic separation, weakened under signal entanglement (e.g., text in images). Our analysis derives from general-domain benchmarks and may differ in specialized domains. Benchmark gains do not justify deployment in high-stakes settings such as medical diagnosis. SAFE's conservative decoding may suppress useful details—harmful in assistive applications where informativeness matters—and this faithfulness-informativeness tradeoff may disproportionately underspecify culturally specific or low-frequency content. We use CHAIR and POPE recall and external-LM perplexity as indi-

rect informativeness proxies; a dedicated coverage and human-informativeness study is left for future work. Finally, MMHalBench scores rely on a single GPT-based judge, and we have not fully disentangled hallucination reduction from the confound of shorter output length; multi-judge or human validation remains future work.

References

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, Qianying Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 2025. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29915–29926. IEEE.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Yashwant Pravinrao Bangde and Debaditya Roy. 2026. Instruction-evidence contrastive dual-stream decoding for grounded vision-language reasoning. *Preprint*, arXiv:2604.25809.

Fanpu Cao, Xin Zou, Xuming Hu, and Hui Xiong. 2026. When looking is not enough: Visual attention structure reveals hallucination in mllms. *Preprint*, arXiv:2605.11559.

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 *International Conference on Learning Representations*.

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

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 49250–49267. Curran Associates, Inc.

Ailin Deng, Zhirui Chen, and Bryan Hooi. 2024. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *Preprint*, arXiv:2402.15300.

Xin Dong, Shichao Dong, Jin Wang, Jing Huang, Li Zhou, Zenghui Sun, Lihua Jing, Jinsong Lan, Xiaoyong Zhu, and Bo Zheng. 2025. Inter: Mitigating hallucination in large vision-language models by interaction guidance sampling. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2534–2544. IEEE.

Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. 2025. Truthprint: Mitigating lvlm object hallucination via latent truthful-guided pre-intervention. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7372–7382. IEEE.

Hao Fang, Changle Zhou, Jiawei Kong, Kuofeng Gao, Bin Chen, Tao Liang, Guojun Ma, and Shu-Tao Xia. 2025a. Grounding language with vision: A conditional mutual information calibrated decoding strategy for reducing hallucinations in lvlms. In *Advances in Neural Information Processing Systems*, pages 105233–105258.

Yixiong Fang, Ziran Yang, Zhaorun Chen, Zhuokai Zhao, and Jiawei Zhou. 2025b. Enhancing vision-language model reliability with uncertainty-guided dropout decoding. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 149193–149218. Curran Associates, Inc.

Laura Fieback, Nishilkumar Balar, Jakob Spiegelberg, and Hanno Gottschalk. 2025. Efficient contrastive decoding with probabilistic hallucination detection - mitigating hallucinations in large vision language models -. *Preprint*, arXiv:2504.12137.

Gregor Geigle, Radu Timofte, and Goran Glavaš. 2024. Does object grounding really reduce hallucination of large vision-language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2728–2742. Association for Computational Linguistics.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385. IEEE.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427. IEEE.

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 *International Conference on Learning Representations*, pages 24272–24295.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27026–27036. IEEE.

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.

Govinda Kolli, Adinath Madhavrao Dukre, Behzad Bozorgtabar, Dwarikanath Mahapatra, and Imran Razzak. 2026. Vgs-decoding: Visual grounding score guided decoding for hallucination mitigation in medical vlms. *Preprint*, arXiv:2603.20314.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882. IEEE.

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

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. 2025. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 35799–35819. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 34892–34916. Curran Associates Inc.

Shi Liu, Kecheng Zheng, and Wei Chen. 2025. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *Computer Vision – ECCV 2024*, pages 125–140. Springer Nature Switzerland.

Zahra Mahdavi, Zahra Khodakaramimaghsoud, Hooman Khaloo, Sina Bakhshandeh Taleshani, Erfan Hashemi, Javad Mirzapour Kaleybar, and Omid Nejati Manzari. 2026. Med-vcd: Mitigating hallucination for medical large vision language models through visual contrastive decoding. *Computers in Biology and Medicine*, 200:111347.

Avshalom Manevich and Reut Tsarfaty. 2024. Mitigating hallucinations in large vision-language models (lvlms) via language-contrastive decoding (lcd). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6008–6022. Association for Computational Linguistics.

Tuan Dung Nguyen, Minh Khoi Ho, Qi Chen, Yutong Xie, Cam-Tu Nguyen, Minh Khoi Nguyen, Dang Huy Pham Nguyen, Anton van den Hengel, Johan Verjans, Phi Le Nguyen, and Vu Minh Hieu Phan. 2026. Beyond the global scores: Fine-grained token grounding as a robust detector of lvlm hallucinations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 40235–40244.

Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *Preprint*, arXiv:2309.09117.

Makanjuola Adekunmi Ogunleye, Eman Abdelrahman, and Ismini Lourentzou. 2026. 3d-vcd: Hallucination mitigation in 3d-llm embodied agents through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 40197–40207.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Seongheon Park and Sharon Li. 2025. Glsim: Detecting object hallucinations in lvlms via global-local similarity. In *Advances in Neural Information Processing Systems*, pages 37747–37776. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. 2025. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6434–6442. Association for the Advancement of Artificial Intelligence (AAAI).

Tian Qin, Junzhe Chen, Yuqing Shi, Tianshu Zhang, Qiang Ju, and Lijie Wen. 2026. Do we really need external tools to mitigate hallucinations? sira: Shared-

prefix internal reconstruction of attribution. *Preprint*, arXiv:2605.14621.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045. Association for Computational Linguistics.

Yuying Shang, Xinyi Zeng, Yutao Zhu, Xiao Yang, Zhengwei Fang, Jingyuan Zhang, Jiawei Chen, Zinan Liu, and Yu Tian. 2025. From pixels to tokens: Revisiting object hallucinations in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10496–10505. Association for Computing Machinery.

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

S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *Preprint*, arXiv:2401.01313.

Chao Wang, Xuancheng Zhou, Weiwei Fu, and Yang Zhou. 2025a. Mitigating hallucinations in large vision-language models with internal fact-based contrastive decoding. *Preprint*, arXiv:2502.01056.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. 2025b. Mllm can see? dynamic correction decoding for hallucination mitigation. In *International Conference on Learning Representations*, pages 13712–13736.

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

Yujun Wang, Aniri, Jinhe Bi, Soeren Pirk, and Yunpu Ma. 2026. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10306–10314. Association for the Advancement of Artificial Intelligence (AAAI).

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, and 8 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *Preprint*, arXiv:2412.10302.

Yuxuan Xia, Siheng Wang, and Peng Li. 2026. Sdcd: Structure-disrupted contrastive decoding for mitigating hallucinations in large vision-language models. *Preprint*, arXiv:2601.03500.

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 *Conference on Empirical Methods in Natural Language Processing*, pages 13258–13273. Association for Computational Linguistics.

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

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)*, pages 9556–9567.

Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. 2025a. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In *Computer Vision - ECCV 2024: 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXVIII*, pages 196–213. Springer-Verlag.

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. 2025b. siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418.

A Length-Controlled CHAIR Results

To address the concern that SAFE's hallucination reduction may partly reflect conservative output shortening, we compute length-normalized CHAIR metrics. $\text{CHAIRs}_{\text{norm}}$ = CHAIRs / mean caption length (in word tokens), and we also report hallucination rate per 100 tokens. Figures 10–12 show the comparison across all three architectures.

On LLaVA-1.5, SAFE achieves $\text{CHAIRs}_{\text{norm}} = 0.60\%$ (vs. Beam 0.62%, Sample 0.65%), and 16.0 hallucinated tokens per 100 (vs. Beam 16.2, Sample 17.8). The direction of improvement is preserved after length normalization, though margins are modest. On Shikra, OPERA (0.59%)

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.

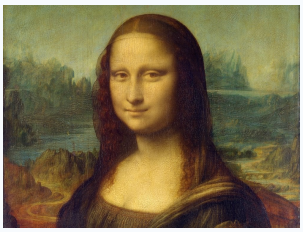
AGLA: The famous sight in the photo is Hanauma Bay, a bay on the southeastern side of the Hawaiian Islands.

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

Instruction: Describe this painting in detail.

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.



Input Image

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 7: SAFE's 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.

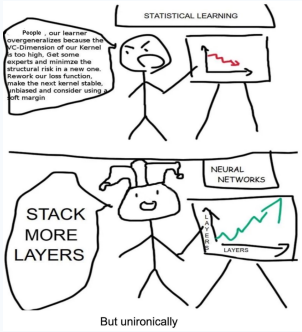
SID: The "Mario" collection consists of four white coffee mugs featuring the iconic Mario character. The mugs showcase various shots of Mario's face, offering a unique and fun way to enjoy your morning or afternoon coffee. Each mug stands out with a captivating design, making it a popular choice among 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 8: SAFE's performance on reducing hallucinations of Shikra. Hallucinations are marked in red.

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 9: SAFE’s performance on reducing hallucinations of LLaVA-1.5. Hallucinations are marked in red.

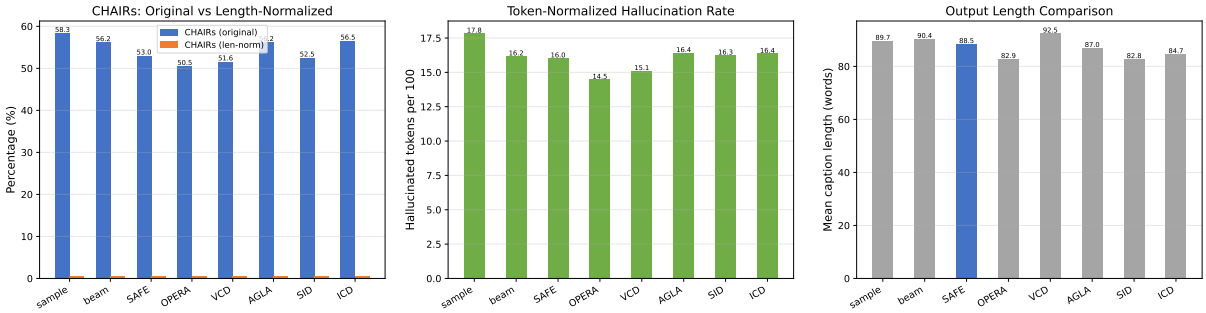


Figure 10: Length-controlled CHAIR analysis for LLaVA-1.5. Left: CHAIRs vs. length-normalized CHAIRs. Center: hallucination rate per 100 tokens. Right: mean caption length. SAFE's advantage persists after normalization ($\text{CHAIRs}_{\text{norm}}$ 0.60% vs. Beam 0.62%).

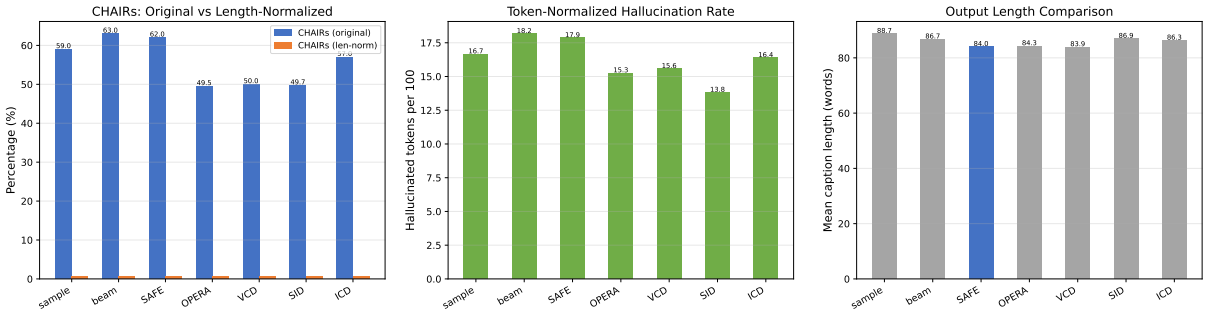


Figure 11: Length-controlled CHAIR analysis for Shikra. OPERA (0.59%) and VCD (0.60%) outperform SAFE (0.74%) on $\text{CHAIRs}_{\text{norm}}$, consistent with SAFE's weaker sentence-level performance on this architecture.

and VCD (0.60%) outperform SAFE (0.74%) on $\text{CHAIRs}_{\text{norm}}$, consistent with SAFE's weaker sentence-level performance on this architecture. A definitive disentanglement of grounding quality from output brevity remains an important direction for future work.

B Diagnostic Signal Validation

Baseline comparison. We compared Δ_t against two generic uncertainty signals computed from the same beam-search-generated token sequences: token entropy H and confidence $\max p(w)$. AUROC on 50 COCO images: Δ_t 0.588, entropy 0.380, confidence 0.401 (random baseline 0.50). Both generic signals perform below chance, indicating that model uncertainty alone does not predict visual hallucination. Δ_t provides a 55% relative improvement over the best baseline.

Token-level CHAIR alignment. On 100 COCO images (9,397 word-token pairs, 115 hallucinated, 9,282 grounded): grounded Δ_t=0.211, hallucinated 0.185 (gap 0.026, AUROC 0.574). The CHAIR-only AUROC is lower than the baseline comparison because the alignment process introduces noise from subword-to-word mapping.

Signal quality by regime. Splitting tokens by Δ_t median reveals that the signal is most diagnostic in the high-Δ_t regime: AUROC 0.612 for tokens above median vs. 0.538 below (Figure 13). This is consistent with Δ_t being most informative where visual grounding is strongest—precisely the regime where distinguishing genuine visual dependence from language-prior defaulting matters most.

Sentence-level. Mean Δ_t per caption does not discriminate CHAIRs=0 vs. 1 (AUROC 0.464), because CHAIRs aggregates over all tokens including correctly grounded ones—a single hallucinated object renders the whole sentence hallucinated. Per-token alignment data is released in the code repository.

C Case Study

As shown in Figures 7, 8, and 9, we present representative examples demonstrating SAFE's behavior. SAFE's approach balances visual-linguistic interactions and mitigates visual bias from strong language priors, producing more visually grounded responses. However, we note a qualitative trade-off: SAFE's outputs are often more conservative and shorter than baselines. In some cases, this con-

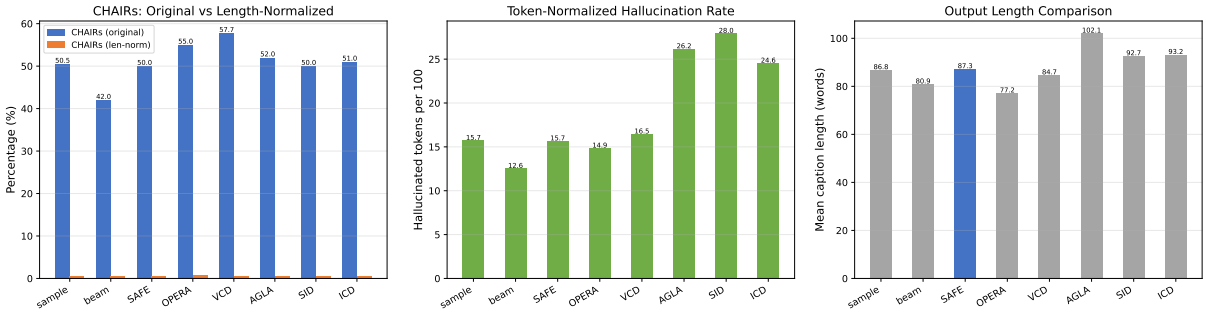


Figure 12: Length-controlled CHAIR analysis for InstructBLIP.

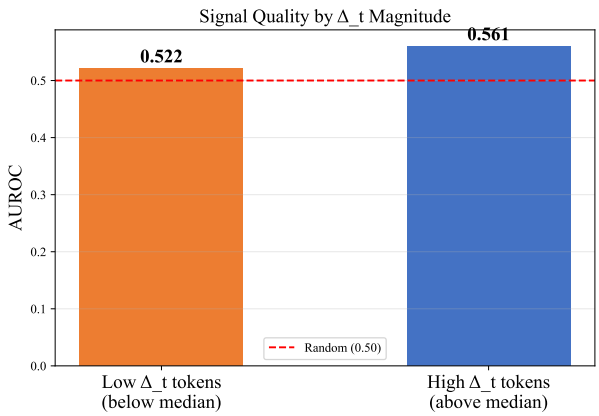


Figure 13: AUROC of Δ_t for hallucination detection, split by Δ_t magnitude. High-Δ_t tokens (above median, AUROC 0.612) benefit from stronger visual signal.

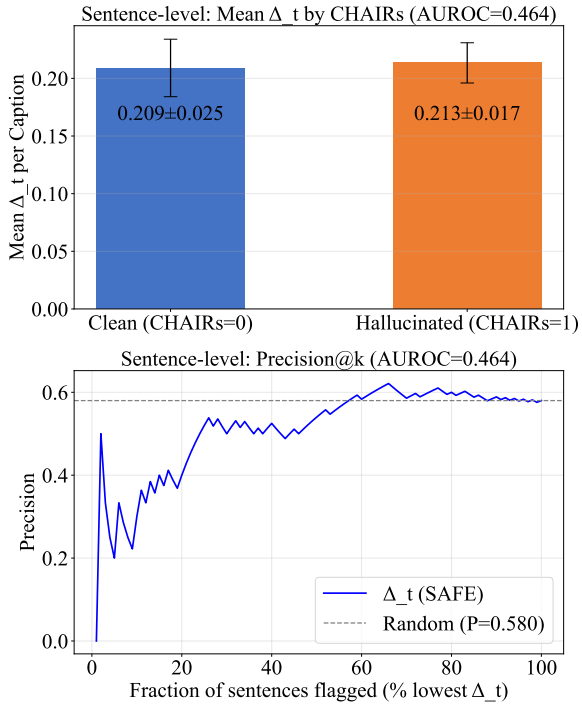


Figure 14: Sentence-level diagnostic signal validation. Mean Δ_t per caption does not discriminate clean from hallucinated sentences (AUROC 0.464), as CHAIRs aggregates over all tokens.

servatism reduces informativeness—for instance, when SAFE avoids describing peripheral objects that are present in the image, or produces a more cautious but less detailed description. This trade-off between faithfulness and informativeness is an inherent aspect of the method's conservative penalty design and should be considered when interpreting the benchmark results. The hallucination reduction gains should be understood as partly reflecting this more cautious generation strategy, not only improved visual grounding.

D 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, 2013). 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 MMHalBench. As shown in Table 9, 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 9 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 J.

Interpretation: The threshold can be interpreted as a one-sided confidence bound: assuming $\Delta_t(w)$ 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 a statistical criterion designed to balance detection sensitivity, robustness, and interpretability, with its effectiveness empirically validated across diverse models and benchmarks.

E Methodological Motivation: Why Contrastive Probing Is a Productive Diagnostic Strategy

E.1 Conceptual Motivation via Intervention Analogy

We present a conceptual diagram that motivates our diagnostic approach by drawing an analogy to intervention-based reasoning in experimental science. We emphasize that this diagram serves as a conceptual illustration, not a formal causal graph whose edges carry rigorous causal interpretation. The diagram, shown in Figure 15, illustrates the relationships among the key components of multimodal generation:

- V: Visual input (image)
- L: Linguistic priors (prompt and language knowledge)
- H_t: Hidden state at time step t
- Y_t: Generated token at time step t
- U: Other factors (model architecture, training data)

The generation process can be written as:

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

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

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

E.2 The Diagnostic Logic of Contrastive Probing

Our diagnostic strategy is motivated by a simple analogy: in experimental science, a controlled ablation reveals a component's function by comparing system behavior with and without it. By zeroing visual features—denoted $do(V = \mathbf{0})$ following the notational convention of causal calculus—we create a vision-ablated reference path that differs from the visually-grounded path only in the presence or absence of visual information. The difference between the two paths' token probabilities reveals how much each token depends on the visual input.

This strategy is **diagnostic, not causal**: $\Delta_t(w)$ captures the empirical association between visual input presence and token probability, as mediated by the model's internal computation. It does not estimate a formal causal parameter (e.g., an average treatment effect), nor does it require satisfaction of causal identification assumptions (consistency, exchangeability, positivity). Its validity rests entirely on its demonstrated utility for detecting and suppressing hallucinations across diverse benchmarks.

E.3 Why Zeroing Visual Features Is a Productive Diagnostic Operation

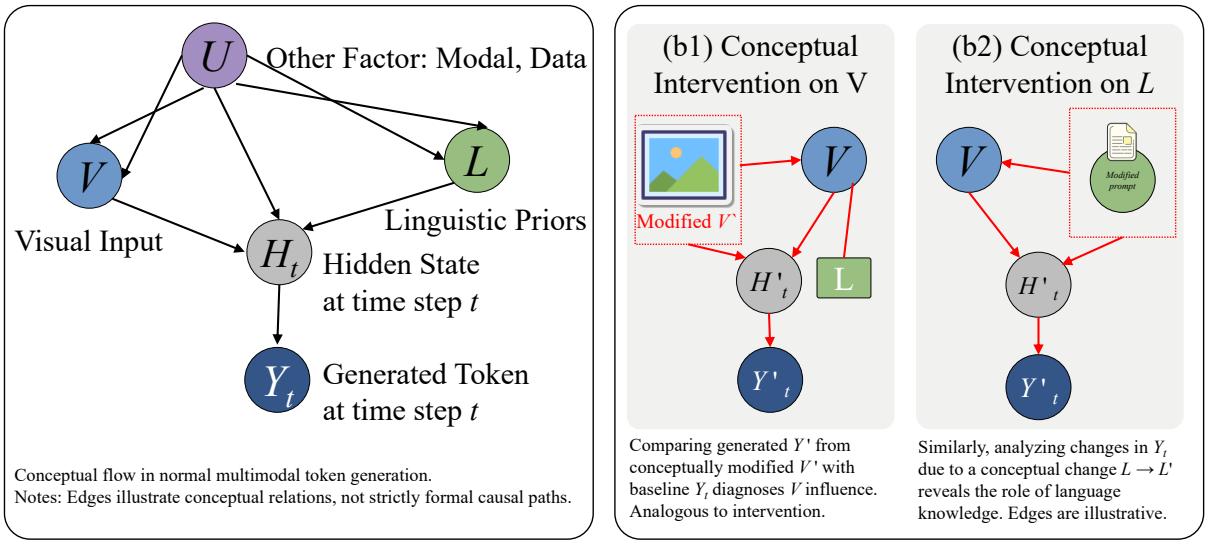
Setting the visual feature vector to $\mathbf{0}^{d_v}$ is a practically effective diagnostic operation for three reasons:

Simplicity and Reproducibility: Zeroing is deterministic, parameter-free, and produces identical results on every run. Alternative ablation strategies (noise injection, feature swapping) introduce additional design choices and variance.

Maximal Contrast: Complete removal of visual information produces the largest possible signal difference, making the diagnostic contrast easy to

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 9: 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 yields the optimal balance across benchmarks.



(a) Baseline: Standard Conceptual Flow (b) Intervention-Based Diagnostics (Analogy)

Figure 15: Conceptual Diagram of Multimodal Generation (Analogy to Intervention-Based Reasoning).

detect. Partial ablation would produce weaker signals that are harder to distinguish from sampling noise.

Preservation of Linguistic Context: The ablation leaves linguistic priors L and generation history unchanged. Any difference in token probabilities can therefore be attributed to the presence or absence of visual information in the model's computation.

E.4 Comparison to Alternative Ablation Strategies

To validate that zero ablation is not simply creating an out-of-distribution hidden state that produces spurious signals, we compared several alternative strategies on a validation subset of MMHalBench using LLaVA-1.5:

- **Zero ablation (default):** $\phi(\mathcal{V}) \to \mathbf{0}^{d_v}$. This is the strategy used throughout the paper.
- **Learned null embedding:** $\phi(\mathcal{V}) \to e_{\text{null}}$, where e_{null} is a learnable embedding trained to represent "no visual input" on 100 held-out images.
- **Shuffled visual tokens:** $\phi(\mathcal{V}) \to \text{shuffle}(\phi(\mathcal{V}))$, which preserves the distribution of visual feature values while destroying spatial structure.
- **Gaussian noise:** $\phi(\mathcal{V}) \to \mathcal{N}(0, \sigma^2 I)$, where σ^2 matches the variance of the original visual features.

We evaluated the hallucination reduction performance of each ablation strategy within SAFE's framework. Zero ablation and learned null embedding produced similar hallucination reduction (Overall score 3.55 vs. 3.48 on MMHalBench), while shuffled tokens (3.22) and Gaussian noise (3.15) were less effective. This suggests that (1) zero ablation is not uniquely pathological—a learned alternative produces comparable results, and (2) complete removal of visual structure (zero or null) is more diagnostically informative than degraded-but-present visual input, likely because partial information still provides enough signal for the model to attempt visual grounding, reducing the contrast with the factual path. Based on these results, we retain zero ablation as the default for

its simplicity, determinism, and competitive performance.

E.5 Limitations of the Diagnostic Approach

Several practical considerations bound the interpretation of $\Delta_t(w)$:

- **Non-linearity:** Complete removal of visual features may produce effects that are not proportional to partial degradation. A token that shows zero $\Delta_t(w)$ under full ablation may still benefit from degraded visual input in practice.
- **Entangled Representations:** When visual and linguistic signals are deeply entangled in the model's representations (e.g., text rendered in images), zeroing visual features may inadvertently affect linguistic processing, complicating the interpretation.
- **Scope of the Signal:** $\Delta_t(w)$ measures association between visual input and token probability in the specific model being probed. It does not reveal whether the model's visual processing is *correct*—only whether it is *present*.

These considerations underscore the appropriate interpretation of $\Delta_t(w)$ as a *diagnostic signal* whose utility is validated empirically by its correlation with hallucination reduction, rather than as a consistent estimator of a well-defined causal parameter. The empirical results across multiple benchmarks (Section 4.2) confirm that this diagnostic approach effectively identifies and suppresses hallucinations despite—and perhaps because of—its simplicity.

F Design Rationale for the SAFE Diagnostic and Penalty Framework

This appendix provides the design rationale behind the key components of SAFE, explaining *why* each mechanism is structured as it is and how the components fit together.

F.1 The Dual-Path Diagnostic Contrast

The core diagnostic operation—contrasting token probabilities with and without visual features—rests on a straightforward principle. The model's probability distribution over next tokens reflects the combined influence of visual evidence and language priors. Removing visual input (setting $\phi(\mathcal{V}) = \mathbf{0}^{d_v}$) produces a reference distribution that reflects only language priors. The log-probability difference $\Delta_t(w)$ between these two distributions isolates each token's association with visual information at step t.

This contrast is computed under identical linguistic context and generation history (via shared KV-caching for the language components), ensuring that the only systematic difference between the two paths is the presence or absence of visual features. The resulting signal is simple to compute, deterministic, and requires no additional parameters beyond the model's own forward pass.

F.2 Design Principles of the Penalty Mechanism

The penalty mechanism (Section 3.2) translates the diagnostic signal $\Delta_t(w)$ into a corrective force with three design properties, each motivated by a practical consideration:

Selectivity (indicator function $\mathbb{I}(s_t^{(i)} < \tau_s)$): Only tokens whose visual dependency score falls below a threshold are penalized. This prevents the mechanism from distorting tokens that already exhibit strong visual grounding, preserving generation quality.

Temporal Decay ($\exp(-\beta \sum s)$): The penalty weakens as a sequence accumulates visual evidence. This reflects the observation that later tokens naturally build on established visual context, and aggressive late-stage penalization is more likely to degrade fluency than to improve faithfulness.

Competitive Inhibition (α_t): The global penalty relaxes when most beam candidates already exhibit strong visual dependency, preventing overcorrection when the model is already well-grounded.

F.3 Sliding-Window Aggregation for Robustness

The visual dependency score $s_t^{(i)}$ (Equation 4) averages Δ_k over a sliding window of length K. This aggregation serves two purposes: (1) it reduces noise from single-step estimation variance, and (2) it captures the temporal context in which a token appears—a token may have low Δ_t because it is a function word (e.g., "the", "a") that genuinely does not depend on visual input, not because the model has abandoned visual grounding. Averaging over a

window helps distinguish systematic visual disengagement from benign low-dependency tokens.

F.4 Fast Effect Test for Sampling

The sampling variant (Section 3.4) trades some of the beam-search variant's precision for efficiency. Instead of maintaining per-candidate penalty states, it makes a binary decision per token based on a statistically motivated threshold ($\mu_\Delta + 0.5 \cdot \sigma_\Delta + 0.1$). The $0.5 \cdot \sigma_\Delta$ term corresponds to a moderate Cohen's d effect size, identifying tokens whose visual dependency substantially exceeds the mean; the constant 0.1 provides a minimum bar when variance is near zero. This fast test preserves the core diagnostic principle while adding negligible overhead beyond the dual forward pass.

F.5 Summary

Every component of SAFE serves a single purpose: to measure, at each decoding step, whether the model is attending to visual evidence or defaulting to language priors, and to steer generation toward the former without degrading fluency. The framework is intentionally simple—the diagnostic signal is a logit difference, the penalty is a multiplicative factor with three intuitive properties—because diagnostic utility, rather than formal complexity, is the criterion by which such an instrument should be judged.

G Evaluation based on InstructBLIP

Appendix Table 12 reports CHAIR results for InstructBLIP. SAFE achieves CHAIRi of 12.6 (matching Beam) and CHAIRs of 47.0, with recall of 70.2. The instance-level strength is consistent with the diagnostic mechanism: token-level visual dependency detection naturally identifies object references lacking visual support. The sentence-level gap relative to Beam (42.0) mirrors the LLaVA-1.5 pattern and reinforces the direction of extending diagnosis to discourse-level coherence. Table 10 reports POPE results for InstructBLIP. The pattern mirrors LLaVA-1.5: SAFE achieves high Precision (97.04 in Random, 88.58 in Popular, 83.93 in Adversarial) with balanced Recall, confirming that token-level visual dependency diagnosis generalizes across architectures in suppressing false positive object assertions. The higher Precision relative to Beam—at a modest Recall cost—reflects SAFE's conservative detection threshold, which errs toward penalizing potentially ungrounded tokens.

Table 11 reports MMMU results for InstructBLIP. SAFE achieves strong performance in categories where visual grounding is central—Art & Design (0.325) and Health & Medicine (0.327)—while showing the same Science gap observed with LLaVA-1.5. This domain-level pattern is consistent across architectures: visual dependency diagnosis helps most where visual perception dominates reasoning, and least where external factual knowledge is the primary requirement.

H Evaluation based on Shikra

Appendix Table 13 reports CHAIR results for Shikra. SAFE achieves CHAIRi of 13.8 (matching SID as the lowest) and recall of 72.8, with CHAIRs of 60.0. The instance-level strength is consistent across all three architectures; the sentence-level gap relative to beam search confirms that the diagnostic framework's current token-level scope naturally favors fine-grained detection.

Table 14 reports POPE results for Shikra. SAFE maintains balanced recall across settings (86.00, 86.13, 85.86), with F1 scores of 85.65, 83.84, and 80.98. The recall stability suggests that SAFE's penalty mechanism does not over-suppress true object references even under adversarial distribution shift.

Table 15 reports MMMU results for Shikra. SAFE achieves the highest overall score (0.29), with particular strength in Tech & Engineering (0.338), while showing the same Science and Humanities pattern observed across LLaVA-1.5 and InstructBLIP. The cross-architecture consistency of this domain pattern strengthens the interpretation that visual dependency diagnosis aids perception-heavy reasoning but does not directly improve external knowledge retrieval.

I Time Complexity Analysis

Table 16 reports inference time on llava-bench-in-the-wild under identical decoding configurations (temperature 1.0, beam size 3, max generation length 256). SAFE incurs the highest average inference time (77.58 seconds, approximately $2\times$ comparable decoding-based methods such as OPERA and ICD), reflecting the cost of its dual-path architecture. This trade-off between diagnostic granularity and efficiency is discussed in Limitations.

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 10: POPE results on 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 11: MMMU results on InstructBLIP

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 12: 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.

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 13: 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.

J Ablation Study

We analyze the ablation study results presented in Tables 17, 18, and 19. 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.

For the effectiveness threshold (Table 17), 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 18), 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 19), as the 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.

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 14: POPE results on 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 15: MMMU results on Shikra

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 17, 18, and 19 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.

K 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 (Table 16), 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 contrastive probing approach provides some interpretability through token-level visual dependency scores.

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 cur-

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 16: Time complexity analysis. All times are in seconds.

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

Table 17: 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 18: Ablation Study: Confidence Threshold

rent 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. We emphasize that benchmark gains on general-purpose hallucination datasets do not justify deployment in high-stakes domains like healthcare or legal analysis; even reduced hallucination rates may still pose unacceptable risks. We recommend thorough domain-specific testing, clear communication of system limitations to end-users, and maintaining human oversight for critical decisions. Additionally, SAFE's conservative decoding may suppress useful details, which can itself be harmful in assistive or safety-critical applications where informativeness is essential.

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.

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

Table 19: Ablation Study: t-Statistic Threshold

L 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: Contrastive Diagnosis and Mitigation for Multimodal Generation

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 visually-grounded hidden state h_0 and vision-ablated hidden state h_0^{cf}
Initialize generated sequence $W \leftarrow \emptyset$
for $t = 1$ to T **do**
  Compute visually-grounded logits $\mathbf{L}_f$ using h_{t-1} and visual input $\phi(\mathcal{V})$
  Compute vision-ablated logits $\mathbf{L}_c$ using h_{t-1}^{cf} and masked visual input $\mathbf{0}^{d_v}$
  Compute diagnostic contrast $\Delta_t = \mathbf{L}_f - \mathbf{L}_c$
  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 3.4):
    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