Latent Attention Denoising: A Training-Free Energy-Based Framework for Mitigating Hallucinations in Vision-Language Models

Zhiwen Luo **Siyu Jiang** **Weilong Jiang** **Kun He***
Huazhong University of Science and Technology
{zhiwenluo, D202481674, librajwl, brooklet60}@hust.edu.cn

Abstract

Visual hallucination remains a major obstacle to the reliability of Large Vision-Language Models (LVLMs). We argue that this issue originates from a fundamental statistical misspecification: the conventional softmax attention implicitly assumes i.i.d. noise, yet real LVLM attention patterns exhibit structured and competitive biases (e.g., attention sinks) that violate this assumption. To address this mismatch, we introduce Latent Attention Denoising (LAD), a principled and training-free framework that recasts attention calibration as a one-step score-based denoising process. LAD employs an interpretable energy function to derive an analytic score and applies a single Langevin-inspired update to actively steer corrupted attention logits toward more faithful configurations. This intervention imposes negligible computational overhead and operates at a speed comparable to standard greedy decoding. Extensive evaluations across diverse architectures confirm that LAD achieves superior performance on both generative and discriminative tasks, effectively mitigating hallucinations while maintaining efficiency comparable to standard decoding.

1 Introduction

Large Vision-Language Models (LVLMs) have demonstrated remarkable capabilities in multimodal perception, reasoning, and generation (Bai et al., 2023; Dai et al., 2023; Ye et al., 2023; Liu et al., 2023; Achiam et al., 2023). Despite these advances, their deployment in real-world settings is significantly hindered by a critical flaw: *visual hallucination* (Zhang et al., 2025b; Huang et al., 2025; Ji et al., 2023; Cui et al., 2023). Models often generate fluent text that contradicts image content, undermining reliability in safety-critical applications such as medical diagnosis (Li et al., 2023a; Xia et al., 2024; Moor et al., 2023) and autonomous navigation (Xu et al., 2024; Sima et al., 2024). Reducing hallucination has therefore become a central challenge for robust multimodal intelligence (Liu et al., 2024b; Rawte et al., 2023; Guan et al., 2024).

Recent analyses suggest that hallucination originates not from insufficient visual representations but from decoder-side *attention sinks* — tokens that disproportionately attract attention and suppress informative visual evidence (Cui et al., 2023; Xiao et al., 2023; Zhou et al., 2024). Existing mitigation strategies generally rely on heuristic adjustments, fail to resolve these competitive biases (Huang et al., 2024a; Liu et al., 2024e; Leng et al., 2024). While effective to some extent, these methods lack a principled foundation and struggle to consistently counteract the competitive biases that drive hallucination.

In this work, we show that attention sinks reflect a deeper issue: a fundamental **statistical misspecification** in the attention mechanism. Softmax attention serves as an optimal estimator only when logit perturbations follow independent and identically distributed (i.i.d.) Gumbel noise (McFadden, 1972). However, LVLMs exhibit structured, non-i.i.d. noise patterns (Xiao et al., 2023; Darcet et al., 2023). These structured biases create a zero-sum competition in which sink tokens consistently suppress visual evidence. As we demonstrate in Section 3, the violation of the i.i.d. assumption renders softmax statistically unreliable for multimodal grounding (Meng and Wang, 2023; Zhai et al., 2023), ultimately driving the model toward language priors and hallucination.

To rectify such statistical misspecification, we necessitate a calibration paradigm that circumvents the restrictive assumptions of the standard softmax estimator. Thereby, we introduce Latent Attention Denoising (LAD), which reformulates attention calibration as a one-step score-based denoising process (Song et al., 2020; Vincent, 2011) (Figure 1). By treating attention as a latent variable cor-

*Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 28748–28777

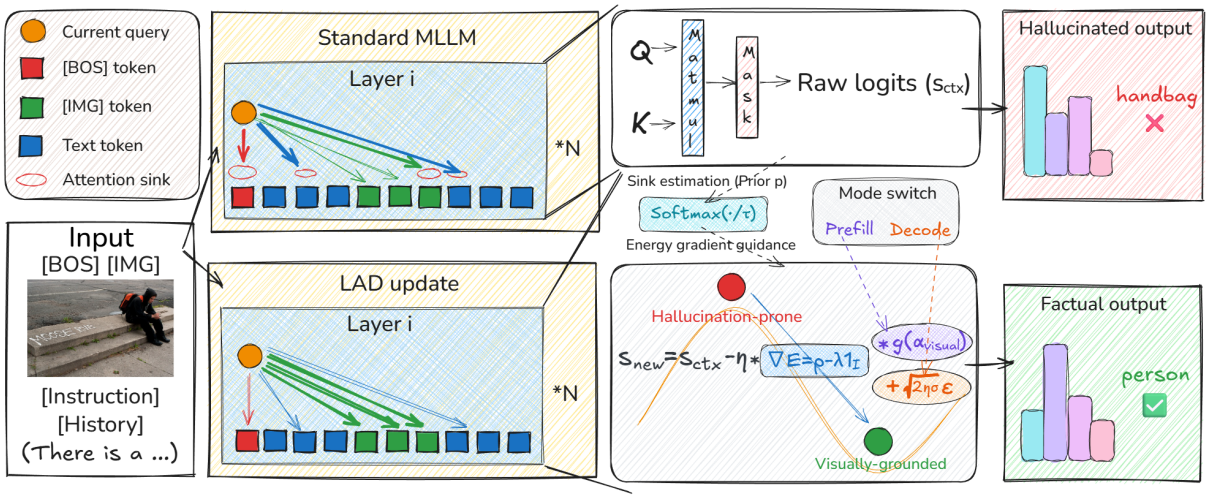


Figure 1: **The Latent Attention Denoising (LAD) Framework. Left:** Standard attention over-allocates probability mass to sinks (red circles), such as special tokens or visual patches. This suppresses relevant image cues (green nodes) and leads to hallucination. **Middle:** LAD infers a sink prior $\mathbf{p}$ and applies **Energy Gradient Guidance** (∇E) to steer the latent state from a high-energy, hallucination-prone region towards a visually-grounded basin. A **Mode Switch** governs updates using a gate $g(\alpha)$ during the Prefill phase and Langevin noise ϵ during the Decode phase. **Right:** The calibrated logits restore attention to valid visual tokens, yielding factual outputs.

rupted by structured noise, we derive a Langevin-inspired update (Welling and Teh, 2011) guided by an interpretable energy function (LeCun et al., 2006) to penalize hallucinatory states. This approach transcends heuristic interventions (Liu et al., 2024e; Huang et al., 2024a), offering a theoretically grounded correction rooted in statistical physics.

Our contributions are threefold:

- We diagnose the **statistical misspecification** of LVLM attention, empirically proving that structured attention sinks violate the i.i.d. Gumbel assumption underlying the softmax operator (McFadden, 1972; Meng and Wang, 2023).
- We propose LAD, a **training-free EBM framework** that calibrates attention via a single-step, analytic Langevin update, incorporating an adaptive gating mechanism to dynamically balance contextual understanding with visual grounding, enabling plug-and-play inference.
- We achieve **state-of-the-art performance** across comprehensive benchmarks and architectures. LAD significantly mitigates hallucinations in diverse LVLMs while incurring negligible computational overhead compared to conventional greedy decoding.

2 Related Work

Hallucination Mitigation via Attention Analysis. Visual hallucination in LVLMs has been extensively documented (Huang et al., 2025; Ji et al., 2023; Bai et al., 2024), with standardized benchmarks like CHAIR (Rohrbach et al., 2018), POPE (Li et al., 2023c), and MME (Fu et al., 2025), as well as recently proposed taxonomies that categorize its diverse failure modes (Cossio, 2025; Chen et al., 2024). Mitigation strategies fall into two classes: training-based methods which rely on fine-tuning or instruction alignment using curated datasets (Liu et al., 2024a; Zhao et al., 2023; Yin et al., 2024; Li et al., 2025; Sun et al., 2024b), and training-free inference interventions (Che et al., 2025; Qian et al., 2025; He et al., 2025), including contrastive decoding (Li et al., 2023b; Zhang et al., 2024b) such as VCD (Leng et al., 2024), ICD (Wang et al., 2024b), and DoLa (Chuang et al., 2023). A prominent line of training-free approaches focuses on the attention mechanism (Chefer et al., 2021; Brauwers and Frasincar, 2021), motivated by the observation of "attention sinks" where specific tokens disproportionately absorb probability mass (Xiao et al., 2023; Kang et al., 2025; Woo et al., 2025).

These interventions typically fall into three categories: (1) Weight reallocation, which boosts visual signals by recycling attention from sinks (Tu et al., 2025; Xie et al., 2025) or directly amplifying im-

age tokens (Liu et al., 2024e; Xu et al., 2025; Zhao et al., 2024); (2) Bias calibration, which mitigates spatial and modality biases (Zhu et al., 2025; Fazli et al., 2025; Zheng and Zhang, 2025) or regulates sink dependencies (Huang et al., 2024a; Zhang et al., 2025a); and (3) Structural constraints, which apply sparse masks (Zhuang et al., 2025; Wang et al., 2023b), manipulate query-key eigenspectra (Tang et al., 2025a), or extract intermediate visual facts (Zhong et al., 2024; Zhou et al., 2025; Tang et al., 2025b) to enforce grounding.

Energy-Based Formulation and Comparisons. Our framework differs from heuristic weight adjustments by adopting a theoretical perspective rooted in Energy-Based Models (EBMs) (LeCun et al., 2006; Xie et al., 2016) and score-based generative modeling (Song et al., 2020; Ho et al., 2020). EBMs have been widely studied for classifier interpretation (Grathwohl et al., 2019; Du and Mordatch, 2019), while score-based methods have been extended to time-series (Yan et al., 2021) and manifold-valued data (De Bortoli et al., 2022). In contrast to these generative settings, we adapt the paradigm for discriminative attention calibration, treating corrupted attention as a latent variable influenced by structured noise.

Rather than modifying weights through ad hoc rules, we formulate the intervention as a single-step Langevin update (Welling and Teh, 2011), derived analytically from the gradient of a defined energy function (Vincent, 2011; Dockhorn et al., 2021). This yields three major distinctions from prior attention-based methods (Liu et al., 2024e; Huang et al., 2024a): it offers a mathematically interpretable update direction grounded in statistical principles; it introduces a novel adaptive gating mechanism that enables mode switching between prefill phase and decode phase; and it establishes a flexible theoretical foundation that naturally accommodates future learning-based or multi-step extensions.

3 Methodology: Energy-Based Attention Calibration

3.1 Preliminaries and Problem Setup

We consider the decoder of a standard Transformer-based LVLM operating on a multimodal input sequence of length N. Let $\mathcal{I} \subset \{1, \ldots, N\}$ denote the indices of **visual tokens** (image patches), while the remaining positions correspond to textual or special-purpose tokens (e.g., system prompts, `[BOS]` sinks).

Each decoder layer employs Multi-Head Attention with H heads. As illustrated in Figure 1, for any head $h \in \{1, \ldots, H\}$, the hidden states $\mathbf{H} \in \mathbb{R}^{N \times d}$ are projected into query, key, and value matrices. The raw attention **logits** for the current query are:

$$\mathbf{s}^{(h)} = \frac{\mathbf{Q}^{(h)}(\mathbf{K}^{(h)})^\top}{\sqrt{d}}, \tag{1}$$

followed by the standard softmax transformation:

$$\mathbf{A}^{(h)} = \text{softmax}(\mathbf{s}^{(h)}). \tag{2}$$

Our method intervenes directly on the latent logits prior to the softmax, allowing us to reshape the attention distribution without modifying the model weights. For clarity, we omit the head superscript (h) in subsequent sections unless cross-head aggregation is explicitly involved.

Standard LVLM inference proceeds in two distinct stages. (1) **Prefill Phase**: the multimodal input (image features and textual instructions) is processed in parallel to initialize the Key-Value (KV) cache. (2) **Decode Phase**: tokens are generated autoregressively. As discussed in Section 3.5, our framework tailors its intervention strategy to balance contextual fidelity and hallucination mitigation.

3.2 Diagnosing the Statistical Misspecification of Softmax

The standard softmax function is rooted in the Random Utility Maximization (RUM) model (McFadden, 1972)which assumes that the utility of token k decomposes as:

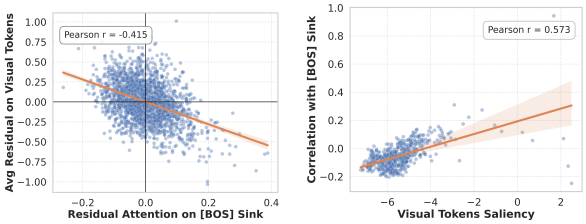
$$U_k = s_k + \varepsilon_k, \quad \varepsilon_k \sim \text{Gumbel}(0, 1). \tag{3}$$

A key requirement is that the noise term ε_k is independent and identically distributed (i.i.d.), implying the Independence of Irrelevant Alternatives (IIA) property:

$$\frac{P(y = i)}{P(y = j)} = \exp(s_i - s_j). \tag{4}$$

Eq. 4 asserts that the relative preference between two tokens should depend *only* on their logits, unaffected by any third-party alternative. A rigorous derivation is provided in **Appendix B.1**.

To examine whether these assumptions hold in LVLMs, we conduct a granular diagnostic analysis



(a) Direct competition (b) Targeted suppression

Figure 2: Diagnosing attention pathologies. **(a)** A strong negative correlation exists between the sink token and visual modality, indicating a zero-sum attentional budget. **(b)** This competition is targeted: the most salient visual tokens (higher average logit) suffer the strongest suppression from the sink.

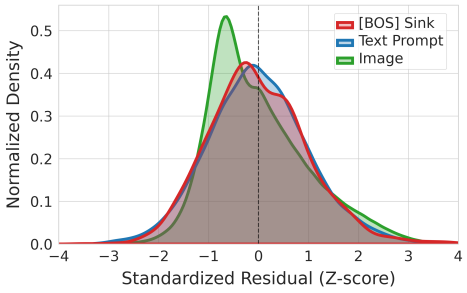


Figure 3: Normalized density of standardized attention residuals.

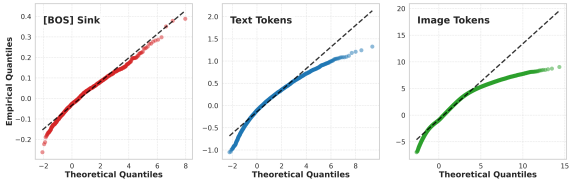


Figure 4: Q-Q plots of attention residuals against a standard Gumbel distribution for [BOS] Sink, Text, and Image tokens.

on LLaVA-1.5 (Liu et al., 2024c). From a typical mid-level decoder layer, we collect attention states over $M = 2,000$ MS-COCO samples (Lin et al., 2014) (see **Appendix A** for experimental protocols). For each token k, we define the *residual logit* as:

$$\bar{s}_k = \frac{1}{M}\sum_{i=1}^{M} s_{i,k}, \quad \varepsilon_{i,k} = s_{i,k} - \bar{s}_k. \tag{5}$$

where $\varepsilon_{i,k}$ for sample i is the deviation from the global mean logit $\bar{s}_k$, acting as the empirical proxy for the theoretical noise term in Eq. 3.

We uncover two pathological behaviors. (1) Zero-sum competition. Sink tokens exhibit strong negative correlation with visual tokens, as illustrated in Figure 2, revealing that sinks actively suppress visual attention rather than acting as passive absorbers. (2) Violation of i.i.d. noise. The empirical residuals for sinks, text tokens, and image tokens follow markedly different distributions (Figures 3–4), contradicting the Gumbel assumption. These results indicate that softmax is statistically misspecified for LVLM attention dynamics (Meng and Wang, 2023), motivating a principled corrective mechanism. See Appendix A.4 for analogous diagnostics on additional architectures.

3.3 Reshaping the Energy Landscape

To rectify this misspecification, we adopt an Energy-Based Model (EBM) perspective (LeCun et al., 2006). We define the probability density of the attention state $\mathbf{A}$ as a Boltzmann distribution $p(\mathbf{A}) \propto \exp(-E(\mathbf{A}))$. Beyond the pretrained energy E_0, we introduce a corrective potential $E_c(\mathbf{A})$ that penalizes hallucinatory patterns, effectively reshaping the energy landscape:

$$E_c(\mathbf{A}) = \underbrace{\mathbf{p}^\top \mathbf{A}}_{\text{Sink Penalty}} - \underbrace{\lambda \mathbf{1}_{\mathcal{I}}^\top \mathbf{A}}_{\text{Visual Incentive}}. \tag{6}$$

Sink Penalty ($\mathbf{p}^\top\mathbf{A}$). The sink prior $\mathbf{p}$ assigns high-energy penalties to tokens consistently acting as sinks across heads and is dynamically estimated at inference time. As illustrated in Figure 5, we hypothesize that true attention sinks act as global attractors across diverse semantic subspaces. To identify them robustly, we aggregate logits across all heads and apply a sharpened softmax:

$$\bar{s}_k = \frac{1}{H}\sum_{h=1}^{H} s_k^{(h)}, \quad p_k = \frac{\exp(\bar{s}_k/\tau)}{\sum_{j=1}^{N}\exp(\bar{s}_j/\tau)}. \tag{7}$$

This head-aggregated formulation captures structural biases independent of semantic context. Sensitivity to temperature parameter τ is studied in **Appendix B.4**.

Visual Incentive ($-\lambda\mathbf{1}_{\mathcal{I}}^\top\mathbf{A}$). Inspired by heuristic visual re-weighting strategies (Liu et al., 2024e; Huang et al., 2024a), we formalize visual grounding as an energy minimization objective. This term assigns negative energy to states that attend to visual tokens (indices $\mathcal{I}$), creating a vector field that "pulls" the attention mass towards image patches.

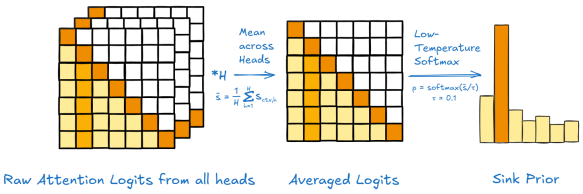


Figure 5: Dynamic Sink Prior estimation.

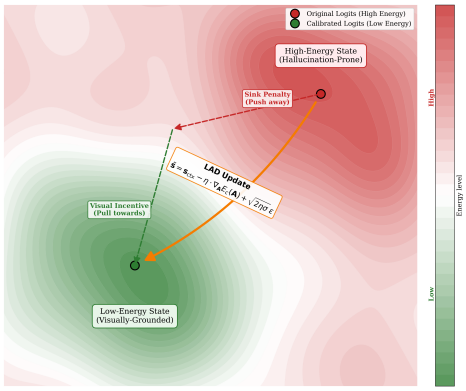


Figure 6: The LAD update viewed as a descent on an energy landscape.

3.4 Langevin-Inspired Logit Calibration

We aim to sample from the corrected distribution $p(\mathbf{A})$. Using Score-Based Generative Modeling (Song et al., 2020), the score function is

$$\nabla_{\mathbf{A}} \log p(\mathbf{A}) = -\nabla_{\mathbf{A}} E_c(\mathbf{A}) = \mathbf{p} - \lambda \mathbf{1}_{\mathcal{I}}. \quad (8)$$

Direct updates in the probability simplex are unstable; therefore, we operate in logit space using a discretized Langevin step (Welling and Teh, 2011):

$$\tilde{\mathbf{s}} = \mathbf{s}_{\text{ctx}} - \eta \nabla_{\mathbf{A}} E_c(\mathbf{A}) + \sqrt{2\eta\sigma}\,\boldsymbol{\epsilon}, \quad (9)$$

where η is the step size (guidance strength), $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is standard Gaussian noise, and σ controls the degree of stochastic exploration. This single-step update effectively "denoises" the attention logits (Figure 6). The deterministic drift ($-\eta\nabla E$) pushes the state away from sinks and towards visual tokens, while stochastic diffusion ($\sqrt{2\eta\sigma}\epsilon$) helps the model to escape spurious minima associated with repetitive loops or rigid language priors. Convergence properties are discussed in **Appendix B.2**.

3.5 Adaptive Gating Across Inference Modes

To avoid over-correcting during pure text processing, we regulate the intervention using a visual affinity score $\alpha_{\text{visual}} = \frac{1}{|\mathcal{I}|}\sum_{j\in\mathcal{I}} s_{\text{ctx},j}$. This score reflects the model's instantaneous demand for visual information and drives a sigmoid gate with bias β and temperature T:

$$g(\alpha_{\text{visual}}) = \sigma\left(\frac{\alpha_{\text{visual}} - \beta}{T}\right). \quad (10)$$

This ensures the correction activates ($g \to 1$) only when the model exhibits endogenous visual alignment. Drawing on the cognitive distinction between ingestion and generation (Kahneman, 2011), we employ a dual-mode strategy.

During the **Prefill Phase** (Understanding Mode), we prioritize linguistic logic via a *gated deterministic* update:

$$\tilde{\mathbf{s}} = \mathbf{s}_{\text{ctx}} - g(\alpha_{\text{visual}}) \cdot \eta \cdot \nabla_{\mathbf{A}} E_c(\mathbf{A}). \quad (11)$$

Conversely, in the **Decode Phase** (Generative Mode), where hallucination risk peaks, we apply the full *stochastic* update (Eq. 9) with $\sigma > 0$. Algorithm 1 summarizes the complete forward pass. Further dynamics and stability analyses are provided in **Appendix B.5**.

3.6 Calibration vs. Suppression

A natural concern is whether reducing sink dominance conflicts with prior observations that attention sinks can stabilize autoregressive decoding in text-only LLMs (Xiao et al., 2023). We view these findings as complementary rather than contradictory. StreamingLLM emphasizes the stabilizing role of sink tokens as anchors for autoregressive score distributions, while a broader literature shows that over-dominant sinks can also divert probability mass from semantically meaningful tokens or grounded visual regions, creating information black holes, repetitive language-prior-driven behaviors, visual attention collapse, and spatial degradation (Sun et al., 2024a; Zhang et al., 2024a; Kang et al., 2025; Darcet et al., 2023; Woo et al., 2025; Huang et al., 2024a; Tu et al., 2025). In LVLMs, this imbalance is especially problematic because it pushes the model toward language-side shortcuts instead of image-grounded reasoning. The issue is therefore not the existence of sinks itself, but their miscalibrated dominance in multimodal contexts.

LAD is designed as a *calibration* mechanism rather than a hard suppression rule. The sink-aware energy in Eq. 6 raises the energy of global attractors, but the update in Eq. 9 only shifts logits continuously and never masks or zeros sink tokens. Moreover, the gated deterministic update in Eq. 11

Algorithm 1 Latent Attention Denoising (LAD) Forward Pass

Inputs: Hidden states $\mathbf{H}$, attention mask $\mathbf{M}$
Hyperparams: $\eta, \sigma, \tau, \beta, T$, image indices $I_{\text{start}}, I_{\text{end}}$

$\mathbf{Q}, \mathbf{K}, \mathbf{V} \leftarrow \text{Project}(\mathbf{H})$
$\mathbf{s}_{\text{raw}} \leftarrow (\mathbf{Q}\mathbf{K}^T)/\sqrt{d_k}$
$\mathbf{s}_{\text{ctx}} \leftarrow \mathbf{s}_{\text{raw}} + \mathbf{M}$
$\bar{\mathbf{s}} \leftarrow \text{Mean}(\mathbf{s}_{\text{ctx}}, \text{dim=heads})$
$\mathbf{p} \leftarrow \text{Softmax}(\bar{\mathbf{s}}/\tau)$
$\nabla_{\mathbf{A}}\mathbf{E} \leftarrow \mathbf{p}$
$\mathbf{m}_{\text{visual}} \leftarrow \text{OneHot}(I_{\text{start}}..I_{\text{end}})$
$\nabla_{\mathbf{A}}\mathbf{E} \leftarrow \nabla_{\mathbf{A}}\mathbf{E} - \mathbf{m}_{\text{visual}}$
$\mathbf{d}_{\text{drift}} \leftarrow -\eta \cdot \nabla_{\mathbf{A}}\mathbf{E}$
if is_generative_mode() **then**
    $\mathbf{Z} \sim \mathcal{N}(0, I)$
    $\mathbf{d}_{\text{diffusion}} \leftarrow \sqrt{2\eta\sigma} \cdot \mathbf{Z}$
    $\tilde{\mathbf{s}} \leftarrow \mathbf{s}_{\text{ctx}} + \mathbf{d}_{\text{drift}} + \mathbf{d}_{\text{diffusion}}$
else
    $\alpha_{\text{visual}} \leftarrow \text{Mean}(\mathbf{s}_{\text{ctx}}[\ldots, \mathcal{I}])$
    $g \leftarrow \text{Sigmoid}((\alpha_{\text{visual}} - \beta)/T)$
    $\mathbf{d}_{\text{gated}} \leftarrow g \cdot \mathbf{d}_{\text{drift}}$
    $\tilde{\mathbf{s}} \leftarrow \mathbf{s}_{\text{ctx}}$
    $\tilde{\mathbf{s}}[\ldots, \mathcal{I}] \leftarrow \tilde{\mathbf{s}}[\ldots, \mathcal{I}] + \mathbf{d}_{\text{gated}}[\ldots, \mathcal{I}]$
end if
$\mathbf{A}_{\text{new}} \leftarrow \text{Softmax}(\tilde{\mathbf{s}})$
$\mathbf{O} \leftarrow \text{Project}(\mathbf{A}_{\text{new}}\mathbf{V})$
return $\mathbf{O}$

attenuates the intervention when the current state is text-dominant, while preserving stronger correction when visual grounding is required. This allows sinks to retain their stabilizing role when they are genuinely helpful, yet prevents them from overwhelming image evidence when grounding should dominate. Additional theoretical discussion of the stability of this soft correction is provided in **Appendix B.2**.

4 Experimental Setup

4.1 Benchmarks and Metrics

We utilize three primary benchmarks that jointly cover both discriminative and generative hallucination scenarios. **POPE** (Li et al., 2023c) evaluates object existence through binary visual question answering (VQA) with random, popular, and adversarial splits designed to probe robustness under distributional shifts. **CHAIR** (Rohrbach et al., 2018) evaluates hallucination in open-ended image captioning on MS-COCO (Lin et al., 2014), using the prompt "*Describe the image in detail.*" **AMBER** (Wang et al., 2023a) provides a holistic assessment across both generative and discriminative tasks, reporting a composite **AMBER Score** that explicitly balances hallucination reduction against reasoning accuracy. To verify that hallucination mitigation does not come at the expense of broader multimodal capability, we additionally report **MME** (Fu et al., 2025), **MM-Vet** (Yu et al., 2024), and **VHTest-YNQ** (Huang et al., 2024b). MME probes perception and cognition, MM-Vet evaluates integrated multimodal reasoning, and VHTest-YNQ measures fine-grained visual hallucination across diverse modes. Together, these benchmarks enable a comprehensive assessment of grounding fidelity, output quality, and general capability preservation. Additional details are provided in Appendix C and Appendix C.3.3.

4.2 Models and Baselines

To evaluate architectural generality, we integrate LAD into four widely-used LVLMs spanning diverse design choices: **LLaVA-1.5-7B** (Liu et al., 2024c), **Shikra-7B** (Chen et al., 2023), **MiniGPT-4-7B** (Zhu et al., 2023), and **InstructBLIP-Vicuna-7B** (Dai et al., 2023). We further extend the evaluation to two recent architectures, **Qwen2-VL-7B** (Wang et al., 2024a) and **LLaVA-NeXT-7B** (Liu et al., 2024d), with detailed results reported in Appendix C.3.2.

We compare against a comprehensive set of baselines, representing the major paradigms in hallucination mitigation: (1) standard decoding strategies, including Greedy decoding, Beam Search (Sutskever et al., 2014), and Nucleus Sampling (Holtzman et al., 2019); (2) contrastive decoding methods, such as ICD (Wang et al., 2024b), VCD (Leng et al., 2024), and SID (Huo et al., 2025); (3) inference-time attention interventions, including OPERA (Huang et al., 2024a) and PAI (Liu et al., 2024e). For all baselines we strictly adhere to the hyperparameter configurations recommended in the original papers or official repositories to ensure a fair and unbiased comparison.

4.3 Implementation Details

LAD is applied to all decoder self-attention blocks during inference. To streamline the hyperparameter search, we fix the visual incentive weight $\lambda = 1.0$, the gating bias $\beta = -2.0$, and the temperature $T = 1.0$. Since the effective correction magnitude is governed by the product $\eta \cdot \lambda$, we only tune the

Table 1: Results on the CHAIR benchmark. C_S and C_I refer to $CHAIR_S$ and $CHAIR_I$. Best and second-best scores are shown in **bold** and underlined, respectively.

	LLaVA-1.5				Shikra				MiniGPT-4				InstructBLIP			
Method	$C_S\downarrow$	$C_I\downarrow$	$F_1\uparrow$	Len.	$C_S\downarrow$	$C_I\downarrow$	$F_1\uparrow$	Len.	$C_S\downarrow$	$C_I\downarrow$	$F_1\uparrow$	Len.	$C_S\downarrow$	$C_I\downarrow$	$F_1\uparrow$	Len.
Greedy	43.0	12.3	77.2	94.5	49.0	14.1	75.3	97.7	36.0	11.9	69.4	84.0	45.4	12.2	74.9	98.0
Beam-5	48.6	13.5	77.8	101.0	49.7	14.3	75.3	100.4	32.6	10.4	70.6	75.6	46.6	13.1	74.1	100.3
Sample	48.4	14.3	73.5	98.6	53.8	15.9	72.4	100.6	36.0	12.2	69.0	85.5	55.8	16.4	69.8	107.7
ICD	44.0	12.0	77.6	97.0	46.2	14.9	72.5	98.3	28.4	9.7	67.5	89.4	51.0	15.8	72.7	108.6
VCD	44.8	12.3	78.1	95.3	49.8	14.5	74.0	96.2	27.4	10.0	**71.1**	70.1	42.4	12.2	75.3	96.0
OPERA	41.4	12.2	**78.4**	91.8	**35.0**	11.5	71.9	69.2	27.8	10.9	68.9	64.3	46.0	14.4	74.2	94.3
SID	45.2	11.7	78.1	94.1	50.4	14.0	74.4	98.0	27.8	10.3	70.3	61.1	42.4	12.5	75.2	100.5
PAI	28.2	7.4	77.3	100.5	36.8	**10.3**	73.9	94.5	22.0	8.5	70.6	56.6	44.2	12.9	74.7	106.9
LAD	**19.6**	**5.1**	77.1	88.5	36.6	**10.3**	**75.9**	92.6	**18.0**	**5.9**	69.8	71.5	**37.0**	**10.2**	**76.2**	108.4

step size η and the noise scale σ.

The specific values are empirically selected based on the sensitivity analysis in Section 5.4, aiming to balance mitigation suppression and caption quality. The configurations used for the main results are: $(\eta, \sigma) = (2.1, 0.1)$ for LLaVA-1.5, $(2.3, 0.1)$ for Shikra, $(0.6, 0.1)$ for MiniGPT-4, and $(0.9, 0.1)$ for InstructBLIP. These settings can be easily adjusted to accommodate different application priorities, such as more aggressive hallucination reduction. All experiments were performed on a single server equipped with eight NVIDIA RTX 3090 GPUs.

5 Results and Analysis

5.1 Image Captioning Performance (CHAIR)

As reported in Table 1, LAD consistently and substantially reduces hallucination rates across all evaluated architectures, while preserving or even improving caption quality (F_1). On LLaVA-1.5, LAD achieves a $CHAIR_S$ of **19.6**, a **55% relative reduction** compared to greedy decoding. Notably, on InstructBLIP, LAD reduces $CHAIR_S$ from 45.4 to **37.0** while simultaneously boosting the F_1 score (74.9 $\rightarrow$ **76.2**). This behavior indicates that the improvement does not stem from conservative truncation or under-generation, but from actively steering the model toward visually grounded content.

5.2 Object Existence Probing (POPE)

Table 2 shows that LAD exhibits strong robustness against object hallucination, particularly on the challenging *Adversarial* split, where spurious linguistic correlations frequently override visual evidence. LAD achieves the highest average F_1 scores on LLaVA-1.5 (**85.71**), MiniGPT-4 (**76.88**), and InstructBLIP (**85.18**). Crucially, these gains on adversarial samples are not accompanied by performance degradation on the Random or Popular splits. This indicates that LAD does not rely on a conservative blind rejection strategy. Instead, it precisely calibrates confidence by breaking the linguistic correlation chains between co-occurring objects, ensuring that generation is grounded in visual reality rather than statistical probability.

Table 2: F_1 scores on POPE across four LVLMs. Best and second-best results are marked in **bold** and underlined. Abbreviations: Random (**Rand.**), Popular (**Pop.**), Adversarial (**Adv.**), Average (**Avg.**).

Method	LLaVA-1.5				Shikra			
	Rand.	Pop.	Adv.	Avg.	Rand.	Pop.	Adv.	Avg.
Greedy	**89.33**	86.39	80.98	85.57	84.16	81.51	78.73	81.47
Beam-5	87.68	85.02	81.67	84.79	84.80	82.36	78.83	82.00
Sample	83.17	81.47	76.95	80.53	82.63	79.65	76.91	79.73
ICD	88.29	86.04	80.49	84.94	83.12	80.41	78.12	80.55
VCD	88.29	85.17	80.30	84.59	81.75	79.90	77.60	79.75
OPERA	88.54	85.97	**82.32**	85.61	85.29	81.29	78.19	81.59
SID	89.04	85.90	81.43	85.46	84.19	81.19	78.75	81.38
PAI	89.27	**86.50**	81.11	85.63	**85.53**	**82.62**	79.29	**82.48**
LAD	88.96	86.48	81.68	**85.71**	84.74	82.29	**79.59**	82.21

Method	MiniGPT-4				InstructBLIP			
	Rand.	Pop.	Adv.	Avg.	Rand.	Pop.	Adv.	Avg.
Greedy	81.15	75.03	72.38	76.19	89.55	83.89	81.07	84.84
Beam-5	70.62	67.00	65.70	67.77	88.65	83.65	80.77	84.36
Sample	59.63	57.24	57.12	58.00	81.68	77.78	76.60	78.69
ICD	76.58	72.40	69.76	72.91	88.65	82.24	80.00	83.63
VCD	76.48	72.82	70.59	73.30	89.12	83.13	81.42	84.56
OPERA	71.63	68.54	66.87	69.01	88.45	83.61	81.66	84.57
SID	81.25	75.13	72.50	76.29	89.05	83.41	80.59	84.35
PAI	80.65	75.42	72.43	76.17	89.32	83.77	80.66	84.58
LAD	**81.74**	**75.89**	**73.01**	**76.88**	**89.79**	**83.94**	**81.81**	**85.18**

5.3 Holistic Evaluation (AMBER)

Table 3 highlights LAD's versatility on the AMBER benchmark, where it achieves the highest composite scores for LLaVA-1.5 (**88.3**), InstructBLIP (**87.9**), and MiniGPT-4 (**79.8**). The framework effectively defies the typical trade-off between generative factuality and discriminative precision. For instance, on MiniGPT-4, LAD reduces the halluci-

Table 3: Comprehensive results on the AMBER benchmark. The AMBER Score balances generative factuality (low CHAIR) and discriminative capability (high F_1). **Bold** and underlined denote the best and second-best results, respectively. Abbreviations: Generative (**Gen.**), Discriminative (**Disc.**), CHAIR (**CH.**), Coverage (**Cov.**), Hallucination rate (**Hal.**), Cogency (**Cog.**), Accuracy (**Acc.**), Precision (**P.**), Recall (**R.**).

Model	Gen. Task				Disc. Task				AMBER Score↑
	CH.↓	Cov.↑	Hal.↓	Cog.↓	Acc.↑	P.↑	R.↑	F_1↑	
LLaVA-1.5									
Greedy	6.1	50.7	27.7	2.9	74.8	**94.5**	65.9	77.6	85.8
Beam-5	7.2	49.3	31.8	3.7	77.0	90.5	**73.0**	80.8	86.8
Sample	10.2	49.8	43.6	3.6	69.5	87.3	63.1	73.2	81.5
ICD	6.4	51.6	31.3	3.3	74.7	94.2	65.9	77.5	85.6
VCD	6.8	51.2	32.8	3.3	74.5	**94.5**	65.3	77.2	85.2
OPERA	6.3	49.3	27.7	3.0	76.1	90.1	71.8	79.9	86.8
SID	5.5	**52.6**	28.4	2.4	**77.6**	93.2	71.4	**80.9**	87.7
PAI	4.5	47.3	21.4	1.6	75.1	**94.5**	66.3	77.9	86.7
LAD	**3.1**	48.2	**17.9**	**1.1**	76.5	93.2	69.6	79.7	**88.3**
Shikra									
Greedy	8.8	51.7	41.4	4.1	74.4	78.9	**83.8**	81.3	86.3
Beam-5	9.6	51.0	42.3	4.3	75.1	80.9	81.6	81.2	85.8
Sample	10.4	50.6	44.8	4.2	73.2	78.4	82.3	80.3	85.0
ICD	9.0	50.8	39.4	3.2	72.4	80.3	77.3	78.8	84.9
VCD	9.2	50.9	38.7	3.4	72.9	77.8	82.8	80.2	85.5
OPERA	7.4	46.9	28.7	**1.9**	74.1	79.3	82.4	80.8	86.7
SID	8.6	**52.3**	40.9	3.0	74.2	79.9	81.7	80.8	86.1
PAI	7.2	50.9	36.1	2.6	**76.9**	**83.1**	81.7	**82.4**	**87.6**
LAD	**6.3**	50.7	**28.3**	**1.9**	75.5	82.8	82.1	81.2	87.5
MiniGPT-4									
Greedy	15.4	**63.3**	65.2	11.1	65.6	90.1	54.0	67.5	76.1
Beam-5	14.5	62.3	60.8	9.7	67.5	81.0	66.6	**73.1**	79.3
Sample	15.4	58.8	62.0	9.0	61.2	70.7	**70.8**	70.7	77.7
ICD	13.0	52.5	46.4	5.1	57.9	67.6	70.1	68.8	77.9
VCD	13.1	59.1	52.6	7.3	65.4	84.8	58.3	69.1	78.0
SID	11.4	58.2	47.3	5.9	65.6	90.1	54.0	67.5	78.1
OPERA	12.3	58.3	49.2	7.2	**67.7**	85.5	61.8	71.7	79.7
PAI	10.7	59.2	42.8	5.6	64.2	90.2	51.7	65.7	77.5
LAD	**7.3**	54.2	**31.5**	**3.5**	65.2	**90.6**	53.0	66.9	**79.8**
InstructBLIP									
Greedy	10.5	50.0	37.3	3.9	72.1	79.3	78.4	78.8	84.2
Beam-5	7.2	52.1	32.2	3.4	76.9	**85.7**	78.2	81.8	87.3
Sample	18.4	52.9	53.9	7.1	76.3	83.9	79.5	81.6	81.6
ICD	7.9	52.6	34.6	4.1	74.5	83.7	76.4	79.9	86.0
VCD	7.0	53.0	33.5	3.2	**77.0**	85.1	79.2	82.0	87.5
OPERA	7.9	52.0	32.1	3.6	76.9	84.6	**79.7**	**82.1**	87.1
SID	6.8	**54.1**	32.3	3.6	75.6	83.0	79.4	81.2	87.2
PAI	6.6	50.1	25.9	3.0	76.6	86.7	76.5	81.3	87.4
LAD	**5.7**	49.7	**25.7**	**2.9**	76.6	85.5	78.0	81.6	**87.9**

nation rate (Hal) by over 50% relative to the greedy baseline (65.2 → **31.5**). Similarly, on InstructBLIP, it secures the lowest hallucination rate among all methods (**25.7**) while maintaining a highly competitive discriminative F_1 (81.6). These results confirm that LAD refines visual grounding precision, simultaneously benefiting open-ended descriptions and rigorous closed-set QA tasks.

5.4 Ablation and Hyperparameter Sensitivity

5.4.1 Component Analysis

Table 4 validates our design on LLaVA-1.5. The **Sink Penalty** proves most critical; its removal spikes $CHAIR_S$ to 35.8, identifying sinks as the primary noise source. Yet, removing the **Visual Incentive** degrades performance to 29.6, confirming the necessity of active visual guidance. Finally, omitting **Adaptive Gating** worsens $CHAIR_S$ to 26.6, underscoring the need for context-aware intervention to preserve linguistic reasoning.

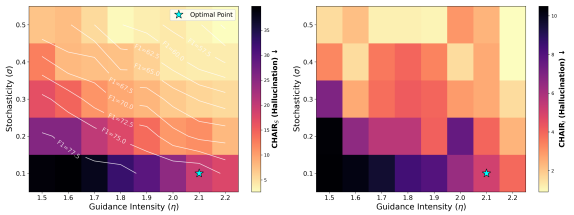


Figure 7: Hyperparameter sensitivity for LLaVA-1.5.

Table 4: Component ablation study of LAD on LLaVA-1.5. Symbols denote: Sink Penalty ($\mathcal{S}$), Visual Incentive ($\mathcal{V}$), Stochasticity (σ), and Adaptive Gating ($\mathcal{G}$). C_S and C_I refer to $CHAIR_S$ and $CHAIR_I$, respectively. The full model uses parameters $\eta = 2.1, \sigma = 0.1$.

Components				Metrics		
$\mathcal{S}$	$\mathcal{V}$	σ	$\mathcal{G}$	$C_S \downarrow$	$C_I \downarrow$	$F_1 \uparrow$
×	✓	✓	✓	35.8	9.4	78.2
✓	×	✓	✓	29.6	7.9	76.0
✓	✓	×	✓	32.8	9.0	78.1
✓	✓	✓	×	26.6	7.4	76.7
✓	✓	✓	✓	**19.6**	**5.1**	**77.1**

5.4.2 Hyperparameter Landscape

Figure 7 reveals a well-defined "basin" of optimal performance. While increasing the guidance strength η suppresses hallucinations, overly aggressive updates degrade F_1 scores. Crucially, stochasticity ($\sigma \approx 0.1$) acts as a regularizer, expanding the optimal η range and enhancing robustness against hyperparameter perturbations. The operating point used in our main experiments ($\eta = 2.1, \sigma = 0.1$) lies well within this stable region. Additional analyses in Appendix D show that this behavior generalizes across architectures.

5.5 Validating the Denoising Hypothesis

To directly validate our core hypothesis, we analyze the distribution of standardized attention logit residuals aggregated across all token categories (including sinks, text, and image tokens), following the identical setup detailed in Appendix A. As illustrated in Figure 8, applying LAD with ($\eta = 2.1, \sigma = 0.1$) dramatically reshapes the heavily skewed baseline residuals (red) into a symmetric

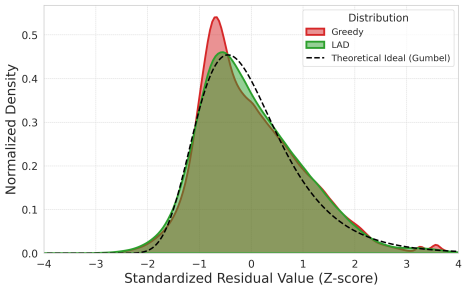


Figure 8: Distributional shift of standardized attention logit residuals aggregated over all token types.

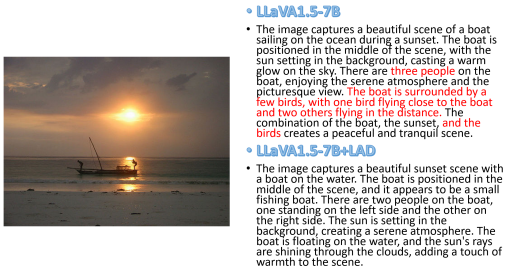


Figure 9: Qualitative comparison. Red text highlights hallucinations by the baseline.

form (green) that closely approximates the theoretical Gumbel ideal. By forcing the latent noise of the global attention landscape to conform to the statistical assumptions required by the softmax operator, LAD restores the validity of the probability estimation, thereby directly mitigating hallucinations stemming from model misspecification.

5.6 Qualitative Analysis

Figure 9 presents representative qualitative examples. While baseline LLaVA-1.5 fabricates "three people" and "birds" due to strong linguistic priors associated with beach scenes, LAD correctly identifies "two people" and suppresses the non-existent birds, demonstrating its ability to override learned language correlations when they conflict with visual evidence. Additional examples are provided in Appendix E.

5.7 Computational Efficiency

Inference-time efficiency is critical for practical deployment. Under the same experimental setting as the CHAIR benchmark (LLaVA-1.5-7B on MS-COCO), we report inference speed and memory usage in Table 5. LAD operates at nearly the same speed as standard greedy decoding (27.99 vs. 30.31 TPS) while being significantly faster than contrastive decoding methods (ICD, VCD) and heavier attention interventions such as OPERA. This efficiency stems from our analytic, single-step update, which introduces negligible overhead ($\mathcal{O}(N)$) compared to the quadratic attention mechanism.

Table 5: Efficiency analysis conducted on LLaVA-1.5.

Method	TPS ↑	Latency (s) ↓	Memory (GB) ↓
Greedy	30.31	3.654	14.02
Beam-5	21.50	5.667	19.57
Sample	30.74	3.832	14.08
ICD	10.53	7.792	15.75
VCD	10.38	7.755	15.74
OPERA	2.50	32.086	21.46
SID	10.58	7.609	15.74
PAI	15.65	9.944	14.88
LAD (ours)	**27.99**	**3.979**	**14.33**

6 Conclusion

We presented Latent Attention Denoising (LAD), a principled framework that reframes attention calibration in LVLMs as a one-step denoising problem. By introducing an interpretable energy formulation and a Langevin-inspired update, LAD explicitly corrects the structured noise and competitive biases inherent in multimodal attention. Extensive evaluations across multiple benchmarks and model architectures demonstrate that our approach consistently and substantially mitigates hallucinations, while incurring negligible computational overhead. Beyond empirical gains, this work offers a statistically grounded perspective on attention reliability, suggesting energy-based calibration as a promising direction for improving the robustness of future Vision-Language Models. We hope this work encourages further exploration of statistically grounded attention modeling in multimodal systems.

Limitations

While LAD establishes a new standard for hallucination mitigation, we acknowledge distinct boundaries in its current application. First, as an inference-time intervention, LAD recalibrates the model's focus but cannot remedy fundamental knowledge deficits or encoder blindness in the pre-trained backbone. Second, our current energy formulation targets object-level and attribute-level inconsistencies; its extension to correcting abstract reasoning fallacies or long-horizon planning errors remains a promising avenue for future research. Finally, while training-free, optimal performance

requires architecture-specific calibration of hyperparameters (η, σ), although our analysis suggests these values remain stable within model families.

Ethical Considerations

LAD significantly improves the factual reliability of LVLMs, potentially broadening their applicability. However, users must remain cognizant that probabilistic corrections do not equate to logical guarantees. No current method, including ours, ensures completely hallucination-free generation. Consequently, the deployment of LAD in high-stakes domains—such as medical diagnostics or autonomous safety systems—should strictly adhere to a human-in-the-loop protocol. We emphasize that enhanced grounding scores should not encourage over-reliance on automated systems without appropriate domain-specific safeguards.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (U22B2017), and the International Cooperation Foundation of Hubei Province, China (2024EHA032).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shun-Ichi Amari. 1998. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Gianni Brauwers and Flavius Frasincar. 2021. A general survey on attention mechanisms in deep learning. *IEEE transactions on knowledge and data engineering*, 35(4):3279–3298.

Liwei Che, Tony Qingze Liu, Jing Jia, Weiyi Qin, Ruixiang Tang, and Vladimir Pavlovic. 2025. Hallucinatory image tokens: A training-free eazy approach to detecting and mitigating object hallucinations in lvlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21635–21644.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791.

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

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. 2024. Multi-object hallucination in vision language models. *Advances in Neural Information Processing Systems*, 37:44393–44418.

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*.

Manuel Cossio. 2025. A comprehensive taxonomy of hallucinations in large language models. *arXiv preprint arXiv:2508.01781*.

Chenhang Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*.

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. *Advances in Neural Information Processing Systems*, 36:49250–49267.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2023. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. 2022. Riemannian score-based generative modelling. *Advances in Neural Information Processing Systems*, 35:2406–2422.

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. 2021. Score-based generative modeling with critically-damped langevin diffusion. *arXiv preprint arXiv:2112.07068*.

Yilun Du and Igor Mordatch. 2019. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 32.

Mehrdad Fazli, Bowen Wei, Ahmet Sari, and Ziwei Zhu. 2025. Mitigating hallucination in large vision-language models via adaptive attention calibration. *arXiv preprint arXiv:2505.21472*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2025. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. 2019. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*.

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

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

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

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

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. 2024b. Visual hallucinations of multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9614–9631, Bangkok, Thailand. Association for Computational Linguistics.

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 *The Thirteenth International Conference on Learning Representations*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*, first paperback edition. Farrar, Straus and Giroux, New York.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations*.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. 2006. A tutorial on energy-based learning. *Predicting structured data*, 1(0).

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2025. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.

Xi-Lin Li. 2018. Preconditioned stochastic gradient descent. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5):1454–1466.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024b. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024d. Llava-next: Improved reasoning, ocr, and world knowledge. LLaVA-NeXT Blog.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

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

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7219–7228.

Daniel McFadden. 1972. Conditional logit analysis of qualitative choice behavior.

Fanfei Meng and Yuxin Wang. 2023. Transformers: Statistical interpretation, architectures and applications. *Authorea Preprints*.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

Jiaye Qian, Ge Zheng, Yuchen Zhu, and Sibei Yang. 2025. Intervene-all-paths: Unified mitigation of LVLM hallucinations across alignment formats. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

H. Risken and T. Frank. 1996. *The Fokker-Planck Equation: Methods of Solution and Applications*. Springer Series in Synergetics. Springer Berlin Heidelberg.

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

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. 2024. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pages 256–274. Springer.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. 2024a. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. 2025a. Intervening anchor token: Decoding strategy in alleviating hallucinations for mllms. In *The Thirteenth International Conference on Learning Representations*.

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, et al. 2025b. Seeing far and clearly: Mitigating hallucinations in mllms with

attention causal decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26147–26159.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Kenneth E Train. 2009. *Discrete choice methods with simulation*. Cambridge university press.

Chongjun Tu, Peng Ye, Dongzhan Zhou, Lei Bai, Gang Yu, Tao Chen, and Wanli Ouyang. 2025. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. *arXiv preprint arXiv:2503.08342*.

Pascal Vincent. 2011. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. 2023a. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023b. Label words are anchors: An information flow perspective for understanding in-context learning. *arXiv preprint arXiv:2305.14160*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *ACL (Findings)*.

Max Welling and Yee W Teh. 2011. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 681–688.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. 2025. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1927–1951.

Peng Xia, Ze Chen, Juanxi Tian, Yangrui Gong, Ruibo Hou, Yue Xu, Zhenbang Wu, Zhiyuan Fan, Yiyang Zhou, Kangyu Zhu, et al. 2024. Cares: A comprehensive benchmark of trustworthiness in medical vision language models. *Advances in Neural Information Processing Systems*, 37:140334–140365.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

Chunzhao Xie, Tongxuan Liu, Lei Jiang, Yuting Zeng, Yunheng Shen, Weizhe Huang, Jing Li, Xiaohua Xu, et al. 2025. Tarac: Mitigating hallucination in lvlms via temporal attention real-time accumulative connection. *arXiv preprint arXiv:2504.04099*.

Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. 2016. A theory of generative convnet. In *International Conference on Machine Learning*, pages 2635–2644. PMLR.

Xinhao Xu, Hui Chen, Mengyao Lyu, Sicheng Zhao, Yizhe Xiong, Zijia Lin, Jungong Han, and Guiguang Ding. 2025. Mitigating hallucinations in multi-modal large language models via image token attention-guided decoding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1571–1590.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. 2024. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*.

Tijin Yan, Hongwei Zhang, Tong Zhou, Yufeng Zhan, and Yuanqing Xia. 2021. Scoregrad: Multivariate probabilistic time series forecasting with continuous energy-based generative models. *arXiv preprint arXiv:2106.10121*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

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

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57730–57754.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. 2024a. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms. *arXiv preprint arXiv:2411.09968*.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. 2025a. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in lvlms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3512–3534.

Yue Zhang, Leyang Cui, Wei Bi, and Shuming Shi. 2024b. Alleviating hallucinations of large language models through induced hallucinations.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2025b. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46.

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

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Haohan Zheng and Zhenguo Zhang. 2025. Modality bias in lvlms: Analyzing and mitigating object hallucination via attention lens. *arXiv preprint arXiv:2508.02419*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. 2024. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11991–12011.

Haoran Zhou, Zihan Zhang, and Hao Chen. 2025. Extracting visual facts from intermediate layers for mitigating hallucinations in multimodal large language models. *arXiv preprint arXiv:2507.15652*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *12th International Conference on Learning Representations, ICLR 2024*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

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

Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. 2025. Vasparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4189–4199.

A Statistical Diagnosis Methodology and Evidence

In the main paper (Section 1), we presented empirical evidence demonstrating the pathological behavior of attention sinks and the violation of the statistical assumptions underlying the softmax function. This appendix details the exact experimental setup, data collection protocols, and mathematical definitions used to derive those diagnostic plots.

A.1 Data Collection and Experimental Setup

To ensure the robustness and reproducibility of our analysis, we fixed all hyperparameters and extraction points. The diagnostic data was collected using **LLaVA-1.5-7B** on the **MS-COCO 2014** validation set.

Configuration. We employed the following specific settings to extract attention logits:

- **Prompt:** We used the standard broad captioning prompt: `["Describe the image in detail."]` to elicit comprehensive attention over the visual features.
- **Sample Size:** We randomly sampled $N = 2000$ image-text pairs from the validation set.
- **Random Seed:** Fixed at `42` to ensure deterministic data selection.
- **Model Depth:** We extracted logits from Decoder Layer **15** (`layer-idx=15`). Mid-to-late layers are empirically known to capture high-level semantic alignments between vision and text, making them ideal for diagnosing hallucination mechanics.
- **Attention Head:** We selected Head **5** (`head-idx=5`) for detailed visualization, though similar trends were observed across other heads.
- **Query Position:** We analyzed the attention distribution originating from the **last token** of the input prompt (`query-idx=-1`). This position is critical as it aggregates information from the entire context (image + instruction) immediately before the model begins generating the first new token.

A.2 Defining Residual Logits

Standard attention analysis often focuses on post-softmax attention weights ($\mathbf{A}$). However, to rigorously test the statistical validity of the softmax operator, we must analyze the pre-softmax logits ($\mathbf{s}$), as the softmax nonlinearity distorts the underlying noise distribution.

As formalized in Eq. 5 of the main text, we define the *residual logit* $\varepsilon_{i,k}$ by subtracting the dataset-wide mean logit $\bar{s}_k$ from the raw score. This subtraction serves a critical statistical purpose: it decouples the deterministic structural bias (e.g., the inherent "popularity" of the `[BOS]` token or specific delimiters) from the stochastic instance-specific noise.

By centering the logits, $\varepsilon_{i,k}$ becomes a direct empirical proxy for the noise term ϵ in the Random Utility Model (RUM). A positive residual ($\varepsilon_{i,k} > 0$) signifies that token k received unexpectedly high attention in a specific sample relative to its baseline behavior, representing genuine semantic signal (or specific noise) rather than global bias. This transformation is a prerequisite for the Q-Q plots and distribution analyses presented in Section 3.2.

A.3 Testing the i.i.d. Gumbel Assumption

The theoretical justification for using the Softmax function as a probability estimator relies on the assumption that the noise terms ε are Independent and Identically Distributed (i.i.d.) according to a **Gumbel(0,1)** distribution. We rigorously tested this hypothesis using three statistical tools:

1. Independence Test via Correlation Analysis (Figure 2). The "Independent" assumption implies that the noise term of the sink token should not influence or correlate with the noise terms of visual tokens. To test this:

- **Direct Competition (Figure 2a):** We computed the Pearson correlation coefficient between the residual logit of the sink token ($\varepsilon_{\text{sink}}$) and the mean residual logit of all visual tokens ($\bar{\varepsilon}_{\text{visual}}$) across the M samples. We visualized this relationship using a scatter plot with a linear regression fit. The strong negative slope (Pearson $r \ll 0$) quantifies the "zero-sum" nature of the attention budget, proving that sink activation directly suppresses visual attention, thereby violating independence.
- **Targeted Suppression (Figure 2b):** We further investigated *which* visual tokens are most affected. For each individual visual token $k \in \mathcal{I}$, we calculated its specific correlation

with the sink: $\rho_k = \text{Corr}(\varepsilon_{\text{sink}}, \varepsilon_k)$. We then plotted these ρ_k values (y-axis) against the token's baseline saliency $\bar{s}_k$ (x-axis). The resulting negative trend indicates that highly salient visual tokens (high $\bar{s}_k$) exhibit the strongest negative correlations (lowest ρ_k) with the sink, confirming that the suppression is structurally targeted rather than random.

2. Distributional Shape Comparison (Kernel Density Estimation). If the "Identically Distributed" assumption holds, the noise distributions for different token types should effectively be the same. We categorized tokens into three groups: `Sink` (e.g., [BOS]), `Text` (instruction tokens), and `Image` (visual tokens). To compare their shapes independent of magnitude, we standardized the residuals within each group. For a residual x belonging to a group with mean μ and standard deviation σ, the Z-score is:

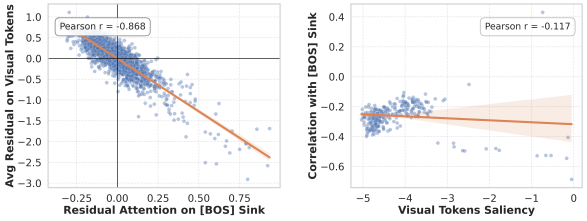
$$z = \frac{x - \mu}{\sigma} \tag{12}$$

We then plotted the Kernel Density Estimates (KDE) of these standardized residuals (Figure 3). The resulting plots show distinct skewness and kurtosis for each group, empirically falsifying the "identical" assumption.

3. Quantile-Quantile (Q-Q) Analysis. To test the specific "Gumbel" assumption, we compared the empirical quantiles of our observed residuals against the theoretical quantiles of the standard Gumbel distribution (Figure 4).

- **X-axis (Theoretical):** The value x such that $P(G < x) = p$, where $G \sim \text{Gumbel}(0, 1)$.
- **Y-axis (Empirical):** The value y such that proportion p of our observed data is less than y.

If the data were perfectly Gumbel-distributed, the points would lie on the $y = x$ identity line (black). The observed heavy deviations, particularly the heavy tails in the Image token distribution, indicate that "extreme" attention events occur far more frequently than the standard Softmax model anticipates. This statistical misspecification suggests that the model lacks the inherent capacity to handle the complex, structured noise found in multimodal contexts.



(a) Direct competition (b) Targeted suppression

Figure 10: Shikra exhibits the same sink-driven competition pattern as LLaVA-1.5. **(a)** The residual attention on the `[BOS]` sink is strongly negatively correlated with the average residual on visual tokens. **(b)** The suppression remains structured rather than random: more salient visual tokens tend to suffer stronger interference from the sink.

A.4 Generalization of Theoretical Motivation

To complement the LLaVA-1.5 analysis in the main text, we provide representative additional-architecture diagnostics generated with the same residual-based analysis pipeline described above. As shown in Figures 10 and 11, the same pathology persists beyond the main-paper case: sink tokens compete directly with visual tokens, and the empirical residual distributions deviate substantially from the i.i.d. Gumbel assumption.

B Theoretical Foundations and Derivations

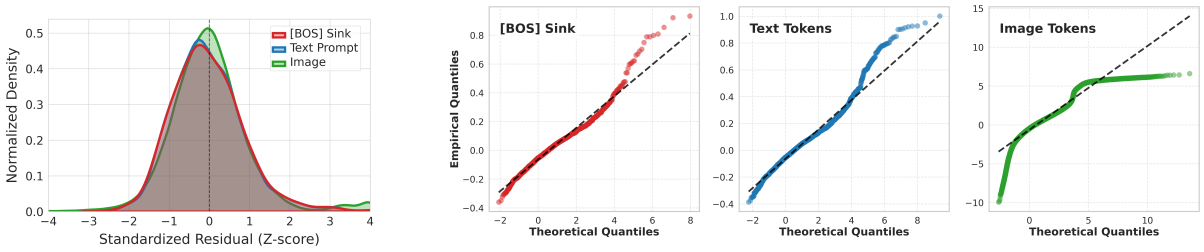
In this section, we provide the rigorous mathematical foundation for the Latent Attention Denoising framework. We begin by deriving the standard softmax from Random Utility Maximization (RUM) theory to explicitly show where the i.i.d. assumption enters and subsequently fails. We then detail the Energy-Based Model (EBM) formulation, the derivation of the score function, and the connection to Langevin Dynamics.

B.1 Derivation of Softmax and the IIA Violation

The softmax function is not merely a heuristic for normalization; it is the exact solution to a specific probabilistic choice problem under strict assumptions.

Random Utility Maximization (RUM). Following the classic discrete choice framework (McFadden, 1972; Train, 2009), let s_k be the deterministic utility (logit) of token k. The perceived utility U_k is modeled as:

$$U_k = s_k + \varepsilon_k \tag{13}$$



(a) Standardized residual density

(b) Q-Q plots against Gumbel(0,1)

Figure 11: Residual distribution diagnostics on Shikra. **(a)** Sink, text, and image tokens follow visibly different standardized densities, rejecting the identically distributed assumption. **(b)** The Q-Q plots further reveal strong departures from the theoretical Gumbel distribution, especially for image tokens, confirming that the softmax noise model remains misspecified beyond the main-paper case.

where ε_k is a random noise term. It is a fundamental result in choice theory that if and only if the error terms ε_k are independent and identically distributed (i.i.d.) following a Gumbel(0, 1) distribution, the probability of selecting the maximal utility token k is given exactly by the softmax function:

$$P(y=k) = P(U_k \geq U_j, \forall j \neq k) = \frac{e^{s_k}}{\sum_{j=1}^{N} e^{s_j}}. \tag{14}$$

Proof of IIA Property. A direct corollary of the i.i.d. assumption is the Independence of Irrelevant Alternatives (IIA). Consider the ratio of probabilities for two options i and j:

$$\frac{P(i)}{P(j)} = \frac{\frac{e^{s_i}}{\sum_k e^{s_k}}}{\frac{e^{s_j}}{\sum_k e^{s_k}}} = \frac{e^{s_i}}{e^{s_j}} = e^{s_i - s_j} \tag{15}$$

This ratio depends *only* on s_i and s_j, implying that the relative preference between two visual tokens should remain invariant regardless of the presence of other tokens.

Failure under Correlation. However, in LVLMs, this independence assumption is violated. If the errors are correlated (e.g., due to attention sinks acting as global suppressors), i.e., $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ where $\Sigma_{ij} \neq 0$, the probability integral does not factorize into the simple softmax form. Instead, $P(i)$ becomes a function of the full covariance matrix:

$$P(i) = \int \mathbb{I}[U_i \geq \max_j U_j] p(\boldsymbol{\varepsilon}; \boldsymbol{\Sigma}) d\boldsymbol{\varepsilon} \tag{16}$$

In this case:

$$\frac{\partial}{\partial s_k}\left(\frac{P(i)}{P(j)}\right) \neq 0 \quad \text{for } k \neq i, j \tag{17}$$

This violation explains why high-probability sink tokens k can disproportionately suppress the ratio between visual tokens i and j, theoretically justifying the need for our corrective framework.

B.2 Langevin Dynamics and The Fokker-Planck Derivation

To sample from the corrected distribution $p_{\text{total}}(\mathbf{A}) \propto \exp(-E_{\text{total}}(\mathbf{A}))$, we employ Langevin Dynamics, a specific realization of score-based generative modeling (Song et al., 2020). We rigorously show that this process converges to the desired Boltzmann distribution.

Stochastic Differential Equation (SDE). We model the time-evolution of the attention state $\mathbf{A}_t$ as a diffusion process governed by the Overdamped Langevin Itô SDE:

$$d\mathbf{A}_t = -\nabla_{\mathbf{A}} E_{\text{total}}(\mathbf{A}_t) dt + \sqrt{2} d\mathbf{W}_t, \tag{18}$$

where $\mathbf{W}_t$ is a standard Brownian motion (Wiener process). This equation describes a particle (the attention distribution) drifting down the energy gradient while being subject to thermal fluctuations.

Convergence via Fokker-Planck. The evolution of the probability density function $\rho(\mathbf{A}, t)$ of the state $\mathbf{A}_t$ is described by the Fokker-Planck equation (Kolmogorov Forward Equation) (Risken and Frank, 1996):

$$\frac{\partial \rho}{\partial t} = \nabla \cdot (\rho \nabla E_{\text{total}}) + \Delta \rho, \tag{19}$$

where $\nabla\cdot$ denotes divergence and Δ is the Laplacian. To prove that the stationary distribution ρ_∞ matches our target Boltzmann distribution $p_{\text{total}}(\mathbf{A}) = \frac{1}{Z} e^{-E_{\text{total}}(\mathbf{A})}$, we verify the condition $\frac{\partial \rho_\infty}{\partial t} = 0$.

Substituting ρ_∞ into the flux term $\mathbf{J} = \rho\nabla E_{\text{total}} + \nabla\rho$:

$$\begin{aligned}\nabla\rho_\infty &= \nabla\left(\frac{1}{Z}e^{-E_{\text{total}}}\right)\\ &= \frac{1}{Z}e^{-E_{\text{total}}}(-\nabla E_{\text{total}})\\ &= -\rho_\infty\nabla E_{\text{total}}. \end{aligned} \tag{20}$$

Substituting this back into the flux equation:

$$\mathbf{J}_\infty = \rho_\infty\nabla E_{\text{total}} + (-\rho_\infty\nabla E_{\text{total}}) = \mathbf{0}. \tag{21}$$

Since the probability flux vanishes, the distribution is stationary. This theoretical result guarantees that, given sufficient time steps, the Langevin update rule drives the attention distribution towards the low-energy, hallucination-free configurations defined by our framework (Welling and Teh, 2011).

B.3 Riemannian Geometry and Logit-Space Mapping

A critical implementation detail is applying the update in the unconstrained logit space $\mathbf{s}$ rather than the simplex Δ^{N-1}. Here, we justify our update rule through the lens of Information Geometry and Riemannian Manifolds (Amari, 1998).

The Geometry of the Simplex. The attention $\mathbf{A}$ resides on a simplex manifold equipped with the Fisher-Rao metric. A naive Euclidean gradient descent on $\mathbf{A}$ is suboptimal. The relationship between infinitesimal changes in logit space $d\mathbf{s}$ and probability space $d\mathbf{A}$ is governed by the Jacobian $\mathcal{J} = \partial\mathbf{A}/\partial\mathbf{s}$:

$$\mathcal{J}_{ij} = A_i(\delta_{ij} - A_j). \tag{22}$$

Strictly applying the chain rule to map the energy gradient $\nabla_{\mathbf{A}}E_{\text{LAD}}$ to logit space would yield:

$$\nabla_{\mathbf{s}}E_{\text{LAD}} = \mathcal{J}^\top\nabla_{\mathbf{A}}E_{\text{LAD}}. \tag{23}$$

However, the term A_i in the Jacobian scales the gradient by the current probability. For visual tokens that have been suppressed to near-zero probability (a common cause of hallucination), $A_i \approx 0$, causing the gradient to vanish ($\nabla_{\mathbf{s}}E \to 0$). This creates a "vanishing gradient" problem where the model cannot recover visual attention once it is lost.

Natural Gradient and Prior Injection. To circumvent this, we adopt a *Preconditioned Langevin Dynamics* approach (Li, 2018). We implicitly apply a preconditioner matrix $\mathbf{M}(\mathbf{A}) \approx \mathcal{J}^{-1}$ (conceptually similar to the inverse Fisher Information Matrix in Natural Gradient Descent) to counteract the curvature of the simplex.

We define the update directly in the logit space as a "force injection":

$$d\mathbf{s}_t = -\nabla_{\mathbf{A}}E_{\text{LAD}}(\mathbf{A}(\mathbf{s}_t))dt + \sqrt{2}d\mathbf{W}_t. \tag{24}$$

By bypassing the Jacobian $\mathcal{J}$, we ensure that the corrective force is constant regardless of the current attention weight A_i. If a visual token i has high corrective potential (low energy), it receives a strong positive boost to s_i even if A_i is currently negligible. This approach can be interpreted as imposing a strong Bayesian prior directly on the latent potentials (logits) rather than on the likelihoods (probabilities), ensuring robust mode recovery.

Discretized Update Rule. Applying the Euler-Maruyama discretization to the logit-space SDE yields our final LAD update rule:

$$\tilde{\mathbf{s}} = \mathbf{s}_{\text{ctx}} - \eta\nabla_{\mathbf{A}}E_{\text{LAD}} + \sqrt{2\eta\sigma}\boldsymbol{\epsilon} \tag{25}$$

$$= \mathbf{s}_{\text{ctx}} - \eta(\mathbf{p} - \lambda\mathbf{1}_{\mathcal{I}}) + \sqrt{2\eta\sigma}\boldsymbol{\epsilon}, \tag{26}$$

where η is the step size and σ is a temperature scaling factor for the noise. This single step approximates the drift towards the high-probability region of the visual-grounded posterior.

B.4 Thermodynamic Analysis of Sink Temperature

The sink prior $\mathbf{p}$ is derived via a Boltzmann distribution over the head-averaged logits $\bar{\mathbf{s}}$, controlled by a temperature parameter τ:

$$p_i(\tau) = \frac{\exp(\bar{s}_i/\tau)}{\sum_{k=1}^{N}\exp(\bar{s}_k/\tau)}. \tag{27}$$

The parameter τ acts as a crucial regulator of the *sparsity* of our sink penalty. We analyze its asymptotic behaviors through the lens of entropy:

High-Temperature Limit ($\tau \to \infty$). As $\tau \to \infty$, the distribution approaches maximum entropy (uniformity):

$$\lim_{\tau\to\infty} p_i = \frac{1}{N}, \quad \forall i. \tag{28}$$

In this regime, the gradient $\nabla E \propto \mathbf{p}$ applies a uniform penalty to all tokens. This is theoretically undesirable as it indiscriminately suppresses valid semantic tokens alongside attention sinks, effectively dampening the global attention magnitude without correcting structural bias.

Low-Temperature Limit ($\tau \to 0$). As $\tau \to 0$, the distribution converges to a hard *argmax* operation:

$$\lim_{\tau \to 0} p_i = \begin{cases} 1 & \text{if } i \in \arg\max_k \bar{s}_k \\ 0 & \text{otherwise} \end{cases} \tag{29}$$

While this perfectly isolates the single most dominant sink (e.g., [BOS]), empirical studies (Kang et al., 2025) suggest that sink behavior is often distributed across a small set of tokens (e.g., delimiters or specific punctuation). A strict one-hot prior would fail to penalize secondary sinks.

Optimal Operating Regime. We select $\tau = 0.1$ to strike a balance between these extremes. This creates a *heavy-tailed* distribution that concentrates mass on the top-k dominant sinks while maintaining differentiability. This "soft-sparsity" ensures that the corrective force is focused on the most pathological tokens while allowing for gradient flow to secondary candidates.

B.5 Dynamics of Adaptive Gating and Stability Analysis

We analyze the adaptive gating mechanism not merely as a heuristic switch, but as a **state-dependent modulation** that ensures system stability. By decomposing the update rule (Eq. 7), we identify a homeostatic control mechanism embedded in the gradient flow.

Gating as Posterior Modulation. The gating function $g(\alpha_{\text{visual}})$ approximates the posterior confidence of visual relevance. Formally, it maps the unbounded logit statistics to a normalized coupling coefficient $[0, 1]$ using a temperature-scaled sigmoid function:

$$\begin{aligned} g(\alpha_{\text{visual}}) &= \sigma\left(\frac{\alpha_{\text{visual}} - \beta}{T}\right) \\ &= \left[1 + \exp\left(-\frac{\alpha_{\text{visual}} - \beta}{T}\right)\right]^{-1} \end{aligned} \tag{30}$$

The hyperparameters serve specific thermodynamic roles:

- **Bias ($\beta = -2.0$):** Calibrated to the sparsity of attention logits, this threshold ensures the gate activates only when the model demonstrates endogenous visual alignment ($\alpha_{\text{visual}} > \beta$), filtering out noise during unimodal processing.
- **Temperature ($T = 1.0$):** This controls the sharpness of the phase transition. Unlike a hard binary switch ($T \to 0$), our choice of $T = 1.0$ maintains a *soft*, differentiable transition. This allows for a smooth, probability-weighted handover between processing modes, preventing gradient discontinuities when the model fluctuates near the decision boundary.

Homeostatic Gradient Dynamics. Substituting the analytic energy gradient, the effective update for a token k is $\Delta s_k = \eta \cdot g(\alpha_{\text{visual}}) \cdot (\mathbf{1}_{k \in \mathcal{I}} \lambda - p_k)$. This reveals two distinct dynamic regimes:

1. Visual Homeostasis (Negative Feedback). For visual tokens ($k \in \mathcal{I}$), the term $(\lambda - p_k)$ acts as a **negative feedback controller**. If attention is insufficient ($p_k \ll \lambda$), the gradient is positive, injecting signal. Crucially, if attention over-saturates ($p_k > \lambda$), the gradient flips negative. This mathematically guarantees that LAD cannot force the distribution to collapse into a single visual patch, maintaining a healthy entropy in the attention distribution.

2. Self-Weighting Sink Suppression. For non-visual tokens ($k \notin \mathcal{I}$), the visual incentive vanishes ($\lambda = 0$), simplifying the update to $\Delta s_k \propto -p_k$. The penalty is thus proportional to the token's pathological severity (its sink probability p_k). Semantic tokens with low p_k remain mathematically invariant, preserving local syntax.

Asymptotic Safety. In the limit of purely textual tasks, the transformer naturally suppresses visual cross-attention, leading to $\alpha_{\text{visual}} \to -\infty$. Consequently, the coupling coefficient decays exponentially:

$$\lim_{\alpha_{\text{visual}} \to -\infty} g(\alpha_{\text{visual}}) = 0 \implies \tilde{\mathbf{s}} \to \mathbf{s}_{\text{ctx}}. \tag{31}$$

This proves that LAD asymptotically reduces to the identity function, safeguarding against the degradation of general language capabilities.

B.6 Algorithmic Complexity Analysis

Computational Complexity. Standard attention requires $\mathcal{O}(N^2 d)$ operations. The LAD overhead consists of:

- Sink estimation: $\mathcal{O}(N)$ (mean across heads).
- Gradient computation: $\mathcal{O}(N)$ (vector subtraction).

- Update: $\mathcal{O}(N)$ (element-wise addition).

Thus, the overhead is $\mathcal{O}(N)$, which is negligible compared to the quadratic attention complexity. This aligns with our empirical efficiency results in Table 5.

C Detailed Experimental Setup and Evaluation Protocols

To ensure reproducibility, this appendix details our experimental framework. We first specify the architectures, inference configurations, and prompt templates for the evaluated LVLMs and baselines. Subsequently, we provide the rigorous definitions and calculation protocols for the POPE (Li et al., 2023c), CHAIR (Rohrbach et al., 2018), and AMBER (Wang et al., 2023a) benchmarks.

C.1 Detailed Experimental Settings

We evaluate four representative Large Vision-Language Models (LVLMs). All models utilize 7B-parameter backbones and were loaded in half-precision (`fp16`) to accommodate the memory constraints of a single NVIDIA GeForce RTX 3090 (24GB VRAM). The specific architectural details and visual token lengths are as follows:

- **LLaVA-1.5-7B:** Built upon the **Llama** (Touvron et al., 2023) architecture. It utilizes a two-layer MLP to project visual features into the LLM's embedding space. With a resolution of 336px, it encodes images into a sequence of **576** visual tokens.

- **Shikra-7B:** Also utilizes the **Llama** architecture. It employs a linear projection layer for vision-language alignment. Unlike LLaVA-1.5, Shikra processes images into a sequence of **256** visual tokens.

- **MiniGPT-4-7B:** Built upon the **Vicuna** (Chiang et al., 2023; Zheng et al., 2023) architecture. It employs a Q-Former and a linear projection layer to align visual features, compressing the visual information into a compact sequence of **32** tokens.

- **InstructBLIP-7B:** Similarly based on the **Vicuna** architecture. It utilizes an instruction-aware Q-Former to extract visual features, also resulting in **32** image tokens.

We strictly adhere to the specific instruction templates required by each model to ensure optimal performance. The templates map the input image (`<ImageHere>`) and the textual query (`<question>`) to the model's input format.

For **LLaVA-1.5** and **Shikra**, we follow established protocols from prior studies such as OPERA (Huang et al., 2024a) and PAI (Liu et al., 2024e) and prepend the system message to define the agent's behavior:

> *System Message:* "A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions."

The exact formatting rules for all evaluated models are listed below:

- **MiniGPT-4:** `###Human: <Img><ImageHere></Img> <question> ###Assistant:`

- **InstructBLIP:** `<ImageHere><question>`

- **Shikra:** `System Message + USER: <im_start><ImageHere><im_end> <question> ASSISTANT:`

- **LLaVA-1.5:** `System Message + USER: <ImageHere> <question> ASSISTANT:`

All experiments were conducted using a fixed random seed of **927** for reproducibility. We set `max-tokens` = 512 for open-ended generative tasks (CHAIR, AMBER-gen) to allow detailed descriptions, and `max-tokens` = 64 for discriminative probes (POPE, AMBER-disc) to ensure concise outputs.

C.2 Metric Definitions

C.2.1 POPE (Polling-based Object Probing Evaluation)

POPE transforms the evaluation of object hallucination into a binary classification task. For a given image-question pair, the model outputs a response $y \in \{\text{“Yes”}, \text{“No”}\}$. Let TP (True Positive) denote cases where the model correctly answers “Yes” for existing objects, TN (True Negative) for correctly answering “No” for non-existing objects, FP (False Positive) for incorrect “Yes”, and FN (False Negative) for incorrect “No”.

We report the standard classification metrics calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{32}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{33}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{34}$$

$$\text{F}_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{35}$$

The F_1-score is the primary metric, as it robustly penalizes both hallucinations (low precision) and object misses (low recall), preventing models from gaming the metric by simply answering "No" to everything.

Table 6 presents the complete evaluation results, including both F_1 scores and Accuracy for all models across Random, Popular, and Adversarial splits.

C.2.2 CHAIR (Captioning Hallucination Assessment with Image Relevance)

CHAIR evaluates the faithfulness of generated captions on the MS-COCO dataset. The computation involves strict text processing steps: tokenization, singularization, and synonym mapping to the 80 MS-COCO object categories using the list provided by Lu et al. (2018). Following the protocol in PAI (Liu et al., 2024e), we randomly sampled 500 images from the validation set using a fixed seed of 927 to ensure consistent comparison.

Let C be the set of all generated captions. For a single caption $c \in C$, let $\mathcal{O}(c)$ be the multiset of object instances mentioned in the caption. Let $\mathcal{O}_{GT}$ be the set of ground-truth objects annotated in the corresponding image. We define the set of hallucinated objects in caption c as:

$$\mathcal{H}(c) = \{o \in \mathcal{O}(c) \mid o \notin \mathcal{O}_{GT}\} \tag{36}$$

We report two variants of the metric:

CHAIR_S (Sentence-level). This measures the percentage of captions that contain at least one hallucinated object:

$$\text{CHAIR}_S = \frac{1}{|C|} \sum_{c \in C} \mathbb{I}\left[|\mathcal{H}(c)| > 0\right] \times 100 \tag{37}$$

where $\mathbb{I}[\cdot]$ is the indicator function.

CHAIR_I (Instance-level). This measures the fraction of hallucinatory object instances relative to the total number of object instances generated:

$$\text{CHAIR}_I = \frac{\sum_{c \in C} |\mathcal{H}(c)|}{\sum_{c \in C} |\mathcal{O}(c)|} \times 100 \tag{38}$$

Control Metrics (F_1 and Len). Solely minimizing CHAIR can lead to trivial solutions where the model generates extremely short or uninformative captions. To monitor the trade-off between hallucination reduction and caption descriptiveness, we report two additional metrics:

- **Len (Average Length):** The average number of words in the generated captions after tokenization. This metric detects if a model reduces hallucinations simply by truncating its outputs.

- **F_1 (Object F_1-score):** The harmonic mean of object-level Precision and Recall. While CHAIR_I captures the error rate (equivalent to $1 - \text{Precision}$), the F_1 score rewards models that not only avoid errors but also successfully recall ground-truth objects. It is defined as:

$$F_1 = 2 \cdot \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{39}$$

where:

$$\begin{aligned} \text{Precision} &= \frac{|\mathcal{O}(c) \cap \mathcal{O}_{GT}|}{|\mathcal{O}(c)|}, \\ \text{Recall} &= \frac{|\mathcal{O}(c) \cap \mathcal{O}_{GT}|}{|\mathcal{O}_{GT}|}. \end{aligned} \tag{40}$$

C.2.3 AMBER (An LLM-free Multi-dimensional Benchmark)

AMBER provides a granular evaluation across both generative and discriminative tasks. We strictly follow the definitions in Wang et al. (2023a).

Generative Task Metrics. Let A_{obj} be the set of annotated ground-truth objects in an image, and H_{obj} be the set of "hallucinatory target objects" (objects not in the image but likely to be hallucinated). For a model response R, we extract the set of generated objects R_{obj}. These are filtered to a refined set R'_{obj} by intersecting with the benchmark's total object vocabulary to remove irrelevant nouns.

We compute four key metrics for the generative task:

1. **CHAIR (Instance Hallucination Rate):** Defined as the proportion of generated objects that are not in the ground truth.

$$\text{CHAIR}(R) = 1 - \frac{|R'_{obj} \cap A_{obj}|}{|R'_{obj}|} \tag{41}$$

2. **Cover (Coverage):** Measures the recall of ground-truth objects.

$$\text{Cover}(R) = \frac{|R'_{obj} \cap A_{obj}|}{|A_{obj}|} \quad (42)$$

3. **Hal (Sentence Hallucination Rate):** A binary metric indicating if the response contains any hallucination.

$$\text{Hal}(R) = \begin{cases} 1 & \text{if CHAIR}(R) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (43)$$

4. **Cog (Cogency):** Measures the model's susceptibility to generating specific "hallucinatory targets" (hard negatives).

$$\text{Cog}(R) = \frac{|R'_{obj} \cap H_{obj}|}{|R'_{obj}|} \quad (44)$$

The final reported values are the averages of these metrics over the entire test set.

Discriminative Task Metrics. This task probes the model with binary questions across three dimensions: **Existence** (Q_{ex}), **Attributes** (Q_{att}), and **Relations** (Q_{rel}). For the aggregated Discriminative Task, we calculate the global confusion matrix across all questions ($Q_{all} = Q_{ex} \cup Q_{att} \cup Q_{rel}$). We report the **Accuracy**, **Precision**, **Recall**, and **F_1-score** derived from this global matrix.

AMBER Score. To facilitate model ranking, we compute the composite AMBER Score, which weighs generative safety and discriminative precision equally. It is defined as:

$$\text{Score} = \frac{(1 - \text{CHAIR}_{avg}) + \text{F}_{1disc}}{2} \times 100 \quad (45)$$

where CHAIR_{avg} is the average instance-level hallucination rate from the generative task, and F_{1disc} is the F_1-score from the discriminative task.

C.3 Additional Experiments

C.3.1 Decoding Strategy Comparison

We further evaluate LAD under three standard decoding strategies: Greedy, Beam-5, and Nucleus Sampling (denoted as *Sample*). Table 7 summarizes the CHAIR and POPE results, while Table 8 reports the full AMBER metrics. Across all four LVLMs, LAD consistently reduces hallucination-related metrics under different search procedures, showing that the gain is not tied to a specific decoding policy.

Table 7: CHAIR and POPE results under different decoding strategies. Sample denotes nucleus sampling. We report CHAIR_S (C_S), CHAIR_I (C_I), CHAIR F_1, and the averaged POPE Accuracy/F_1. Within each decoding pair, the better result is marked in **bold**. Lower is better for C_S/C_I; higher is better for the remaining metrics.

Method	CHAIR			POPE	
	$C_S\downarrow$	$C_I\downarrow$	$F_1\uparrow$	Avg. Acc.↑	Avg. $F_1\uparrow$
LLaVA-1.5					
Greedy	43.0	12.3	**77.2**	84.83	85.57
+LAD	**19.6**	**5.1**	77.1	**85.39**	**85.71**
Beam-5	48.6	13.5	77.8	85.04	84.79
+LAD	**25.2**	**5.5**	**78.1**	**85.63**	**85.77**
Sample	48.4	14.3	**73.5**	79.39	80.53
+LAD	35.8	9.9	73.1	80.52	81.11
Shikra					
Greedy	49.0	14.1	75.3	80.09	81.47
+LAD	**36.6**	**10.3**	**75.9**	**81.98**	**82.21**
Beam-5	49.7	14.3	75.3	80.89	82.00
+LAD	**32.8**	**9.4**	**76.1**	**83.20**	**83.23**
Sample	53.8	15.9	**72.4**	78.28	79.73
+LAD	**39.6**	**11.2**	**72.4**	**81.20**	**81.13**
MiniGPT-4					
Greedy	36.0	11.9	69.4	76.48	76.19
+LAD	**18.0**	**5.9**	**69.8**	**77.69**	**76.88**
Beam-5	32.6	10.4	**70.6**	**72.54**	67.77
+LAD	**20.8**	**7.5**	70.3	70.85	**69.91**
Sample	36.0	12.2	69.0	57.39	58.00
+LAD	**25.6**	**9.5**	**69.2**	**58.37**	**60.65**
InstructBLIP					
Greedy	45.4	12.2	74.9	84.05	84.84
+LAD	**37.0**	**10.2**	**76.2**	**84.71**	**85.18**
Beam-5	46.6	13.1	74.1	**84.32**	84.36
+LAD	**41.4**	**10.7**	**76.3**	84.13	**84.48**
Sample	55.8	16.4	69.8	77.23	78.69
+LAD	**50.4**	**12.9**	**72.8**	**78.33**	**80.08**

C.3.2 Results on Modern Architectures

We further extend the evaluation to two recent LVLMs, Qwen2-VL (Wang et al., 2024a) and LLaVA-NeXT (Liu et al., 2024d), to assess whether the gains of LAD persist on stronger modern architectures. Both models use 7B-parameter backbones and were evaluated in half-precision (`fp16`) under the same inference protocol as the original four models. For LAD, we use $(\eta, \sigma) = (0.3, 0.1)$. Their architectural characteristics are summarized below:

- **Qwen2-VL-7B:** Built upon the **Qwen2** archi-

Table 6: Comprehensive POPE results (F_1 and Accuracy) across four LVLMs. This table supplements the summarized F_1 results in the main text. LAD consistently achieves robust performance across both metrics, balancing generative diversity with factual accuracy. Abbreviations: Random (**Rand.**), Popular (**Pop.**), Adversarial (**Adv.**), Average (**Avg.**).

Model	Method	Random		Popular		Adversarial		Average	
		F_1 ↑	Acc. ↑	F_1 ↑	Acc. ↑	F_1 ↑	Acc. ↑	**F_1 ↑**	**Acc. ↑**
LLaVA-1.5	Greedy	**89.33**	89.37	86.39	86.00	80.98	79.13	85.57	84.83
	Beam-5	87.68	88.33	85.02	85.40	81.67	81.40	84.79	85.04
	Sample	83.17	82.90	81.47	80.73	76.95	74.53	80.53	79.39
	ICD	88.29	88.30	86.04	85.70	80.49	78.63	84.94	84.21
	VCD	88.29	88.23	85.17	84.57	80.30	78.27	84.59	83.69
	OPERA	88.54	88.69	85.97	86.17	**82.32**	**81.80**	85.61	**85.55**
	SID	89.04	**89.40**	85.90	85.90	81.43	80.40	85.46	85.23
	PAI	89.27	89.30	**86.50**	86.13	81.11	79.30	85.63	84.91
	LAD	88.96	89.20	86.48	**86.43**	81.68	80.53	**85.71**	85.39
Shikra	Greedy	84.16	83.57	81.51	80.23	78.73	76.47	81.47	80.09
	Beam-5	84.80	84.43	82.36	81.47	78.83	76.77	82.00	80.89
	Sample	82.63	82.00	79.65	78.43	76.91	74.40	79.73	78.28
	ICD	83.12	82.57	80.41	79.13	78.12	76.00	80.55	79.23
	VCD	81.75	81.13	79.90	78.77	77.60	75.63	79.75	78.51
	SID	84.19	83.80	81.19	80.03	78.75	76.73	81.38	80.19
	OPERA	85.29	85.29	81.29	79.90	78.19	75.63	81.59	80.00
	PAI	**85.53**	**86.07**	**82.62**	**82.77**	79.29	78.60	**82.48**	**82.48**
	LAD	84.74	85.07	82.29	82.20	**79.59**	**78.73**	82.21	81.98
MiniGPT-4	Greedy	81.15	82.73	75.03	75.17	72.38	71.53	76.19	76.48
	Beam-5	70.62	76.07	67.00	71.63	65.70	69.93	67.77	72.54
	Sample	59.63	60.33	57.24	56.03	57.12	55.80	58.00	57.39
	ICD	76.58	78.93	72.40	73.60	69.75	69.97	72.91	74.17
	VCD	76.48	79.13	72.82	74.50	70.59	71.53	73.30	75.05
	SID	81.25	**82.80**	75.13	75.23	72.50	71.63	76.29	76.55
	OPERA	71.63	76.15	68.54	**76.15**	66.87	71.07	69.01	73.47
	PAI	80.65	82.23	75.42	75.80	72.43	71.73	76.17	76.59
	LAD	**81.74**	82.37	**75.89**	75.89	**73.01**	**74.80**	**76.88**	**77.69**
InstructBLIP	Greedy	89.55	89.67	83.89	83.07	81.07	79.40	84.84	84.05
	Beam-5	88.65	89.27	83.65	83.63	80.77	80.07	84.36	84.32
	Sample	81.68	81.40	77.78	76.00	76.60	74.30	78.69	77.23
	ICD	88.65	88.57	82.24	80.80	80.00	77.77	83.63	82.38
	VCD	89.12	89.53	83.13	82.63	81.42	80.47	84.56	84.21
	OPERA	88.45	88.97	83.61	**83.93**	81.66	**81.60**	84.57	**84.83**
	SID	89.05	88.97	83.41	82.17	80.59	78.43	84.35	83.19
	PAI	89.32	89.40	83.77	82.87	80.66	78.80	84.58	83.69
	LAD	**89.79**	**90.10**	**83.94**	83.37	**81.81**	80.67	**85.18**	84.71

tecture. It adopts a native dynamic-resolution visual encoder together with Multimodal Rotary Position Embedding (M-RoPE), allowing the model to preserve fine-grained details across images of varying aspect ratios and resolutions. Consequently, the number of visual tokens is **input-adaptive** rather than fixed.

- **LLaVA-NeXT-7B:** An improved **LLaVA**-family model that retains the minimalist connector design of LLaVA-1.5 while upgrading image encoding with the **AnyRes** strategy. It supports higher-resolution inputs with up to **4× more pixels** (e.g., up to 672×672 or elongated high-resolution layouts), enabling substantially richer visual detail than LLaVA-1.5.

Table 9 reports a compact summary using $CHAIR_S$, $CHAIR_I$, averaged POPE Accuracy/F_1, and AMBER Score. The improvements remain consistent across all decoding strategies, indicating that LAD transfers smoothly to newer high-capacity models.

Table 8: Comprehensive AMBER results under different decoding strategies. Sample denotes nucleus sampling. We report generative and discriminative metrics together with the overall AMBER Score. Within each decoding pair, the better result is marked in **bold**. Abbreviations: Generative (**Gen.**), Discriminative (**Disc.**), Coverage (**Cov.**), Hallucination rate (**Hal.**), Cogency (**Cog.**), and Accuracy (**Acc.**).

Method	Gen. Task				Disc. Task		AMBER Score↑
	CH.↓	Cov.↑	Hal.↓	Cog.↓	Acc.↑	F_1↑	
LLaVA-1.5							
Greedy	6.1	**50.7**	27.7	2.9	74.8	77.6	85.8
+LAD	**3.1**	48.2	**17.9**	**1.1**	**76.5**	**79.7**	**88.3**
Beam-5	7.2	**49.3**	31.8	3.7	77.0	80.8	86.8
+LAD	**5.3**	48.7	**22.5**	**2.5**	**77.1**	**81.0**	**87.9**
Sample	10.2	49.8	43.6	3.6	69.5	73.2	81.5
+LAD	**9.1**	**51.7**	**43.5**	**2.6**	**69.9**	**74.2**	**82.6**
Shikra							
Greedy	8.8	**51.7**	41.4	4.1	74.4	**81.3**	86.3
+LAD	**6.3**	50.7	**28.3**	**1.9**	**75.5**	81.2	**87.5**
Beam-5	9.6	**51.0**	42.3	4.3	75.1	81.2	85.8
+LAD	**6.8**	49.8	**30.5**	**2.4**	**76.7**	**82.3**	**87.8**
Sample	10.4	**50.6**	44.8	4.2	73.2	80.3	85.0
+LAD	**8.2**	50.5	**37.9**	**2.5**	**75.0**	**81.0**	**86.4**
MiniGPT-4							
Greedy	15.4	**63.3**	65.2	11.1	**65.6**	**67.5**	76.1
+LAD	**7.3**	54.2	**31.5**	**3.5**	64.5	65.9	**79.3**
Beam-5	14.5	**62.3**	60.8	9.7	**67.5**	**73.1**	79.3
+LAD	**8.5**	59.6	**32.3**	**3.5**	66.5	71.6	**81.6**
Sample	15.4	**58.8**	62.0	9.0	**61.2**	**70.7**	77.7
+LAD	**11.3**	56.9	**38.4**	**4.3**	60.2	69.6	**79.2**
InstructBLIP							
Greedy	7.3	**53.7**	33.8	3.5	75.5	80.5	86.6
+LAD	**6.3**	53.6	**31.0**	**3.0**	**76.6**	**81.6**	**87.7**
Beam-5	7.5	**52.6**	34.9	3.4	74.7	79.4	86.0
+LAD	**6.2**	51.6	**30.0**	**2.9**	**74.8**	**79.9**	**86.9**
Sample	11.8	**53.0**	49.1	4.8	68.1	74.2	81.2
+LAD	**9.8**	52.6	**45.7**	**4.3**	**69.7**	**76.4**	**83.3**

C.3.3 General Benchmark Results on Six Models

We further consolidate the general capability evaluation across all six LVLMs using three widely adopted benchmarks: **MME** (Fu et al., 2025), **MM-Vet** (Yu et al., 2024), and **VHTest-YNQ** (Huang et al., 2024b). Following the protocol summarized in our rebuttal, MME covers 14 perception and cognition subtasks such as OCR, position recognition, counting, and color; MM-Vet evaluates integrated capabilities including recognition, knowledge, OCR, spatial awareness, language generation, and math; and VHTest targets eight fine-grained hallucination modes, for which we adopt the YNQ benchmark for automated evaluation.

Table 9: Results on two modern LVLMs under different decoding strategies. Sample denotes nucleus sampling. We report $CHAIR_S$ (C_S), $CHAIR_I$ (C_I), averaged POPE Accuracy/F_1, and AMBER Score. Within each decoding pair, the better result is marked in **bold**. Lower is better for C_S/C_I; higher is better for the remaining metrics.

Method	CHAIR		POPE		AMBER Score↑
	C_S↓	C_I↓	Avg. Acc.↑	Avg. F_1↑	
Qwen2-VL					
Greedy	29.2	9.9	88.09	87.96	91.4
+LAD	**26.6**	**6.3**	**89.39**	**89.08**	**91.9**
Beam-5	33.8	6.9	88.70	88.00	91.3
+LAD	**27.8**	**6.5**	**89.03**	**88.41**	**92.6**
Sample	32.4	8.6	89.01	88.70	91.5
+LAD	**27.2**	**7.3**	**89.39**	**89.17**	**92.2**
LLaVA-NeXT					
Greedy	32.4	8.1	88.84	88.79	87.6
+LAD	**25.6**	**7.5**	**88.92**	**88.86**	**88.0**
Beam-5	30.6	8.6	88.55	88.21	88.2
+LAD	**25.0**	**7.7**	**88.77**	**88.47**	**88.9**
Sample	36.6	10.4	86.26	86.04	84.9
+LAD	**31.8**	**9.0**	**86.81**	**86.61**	**85.8**

MME (Total). Each MME subtask contains paired yes/no questions for every image. Let $z_i^{(1)}, z_i^{(2)} \in \{0, 1\}$ denote the correctness indicators for the two questions associated with image i. The official subtask metrics are:

$$\text{Acc} = \frac{1}{2N} \sum_{i=1}^{N} \left(z_i^{(1)} + z_i^{(2)} \right) \times 100 \tag{46}$$

$$\text{Acc}^{+} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I} \left[z_i^{(1)} = 1 \wedge z_i^{(2)} = 1 \right] \times 100 \tag{47}$$

where N is the number of images in the subtask. The reported **MME (Total)** score is the sum of $(\text{Acc} + \text{Acc}^{+})$ over all 14 subtasks.

MM-Vet. MM-Vet uses an official LLM-based evaluator to score each open-ended prediction against the reference answer. Let $s_i \in [0, 1]$ denote the normalized score assigned to the i-th response. We report the official overall score:

$$\text{MM-Vet} = \frac{100}{N} \sum_{i=1}^{N} s_i \tag{48}$$

where N is the total number of benchmark instances.

Table 10: General benchmark results on six LVLMs. We report MME (Total), MM-Vet, and VHTest-YNQ (Overall). Within each Greedy pair, the better result is marked in **bold**. Higher is better for all metrics.

Model	Method	MME↑	MM-Vet↑	VHTest-YNQ↑
LLaVA-1.5	Greedy	1752.62	**31.9**	56.75
	+LAD	**1775.22**	30.2	**56.83**
Shikra	Greedy	893.02	24.4	52.42
	+LAD	**899.64**	**25.3**	**52.58**
InstructBLIP	Greedy	1197.94	25.8	52.42
	+LAD	**1200.41**	**26.1**	**53.17**
MiniGPT-4	Greedy	**1073.50**	**22.6**	49.92
	+LAD	1045.46	19.1	**50.42**
Qwen2-VL	Greedy	2287.08	56.6	67.17
	+LAD	**2300.10**	**56.7**	**67.67**
LLaVA-NeXT	Greedy	1721.25	43.8	58.00
	+LAD	**1748.86**	**44.4**	**58.50**

VHTest-YNQ (Overall). VHTest contains 1,200 visual-hallucination instances spanning eight modes: existence, shape, color, orientation, OCR, size, position, and counting. Under the adopted YNQ setting, each instance is evaluated as a binary yes/no question. Let $c_i \in \{0, 1\}$ denote whether the prediction exactly matches the ground-truth answer for instance i. We report the overall YNQ accuracy:

$$\text{VHTest-YNQ} = \frac{1}{N} \sum_{i=1}^{N} c_i \times 100 \qquad (49)$$

Table 10 compares the greedy baseline and its LAD-augmented counterpart for each model. Overall, LAD preserves or improves the general benchmark performance while maintaining the gains in hallucination mitigation.

D Additional Ablation Studies

Here we provide the hyperparameter sensitivity analyses for Shikra, MiniGPT-4 and InstructBLIP, complementing the LLaVA-1.5 analysis in the main text.

Figures 12, 13, and 14 illustrate the energy landscapes for each architecture. Consistent with the LLaVA-1.5 results, all models exhibit a distinct "basin of optimality" where hallucination is significantly suppressed without degrading caption quality (F_1). However, the topology varies by alignment paradigm: models utilizing **dense visual projection** (LLaVA-1.5, Shikra) tolerate stronger guidance ($\eta \in [2.1, 2.3]$), whereas models employing **Q-Former compression** (MiniGPT-4, InstructBLIP) feature sharper landscapes requiring subtler intervention ($\eta \in [0.6, 0.9]$) to avoid disrupting the highly concentrated visual signals. Detailed interpretations are provided in the respective figure captions.

E Additional Qualitative Examples

In this section, we present an extended gallery of qualitative comparisons to substantiate the effectiveness of LAD across diverse architectures. Figures 15, 16, 17, and 18 showcase side-by-side generations from LLaVA-1.5, Shikra, MiniGPT-4, and InstructBLIP, respectively. These examples highlight specific failure modes of the baseline models—ranging from object fabrication and incorrect counting to contextual hallucination—and demonstrate how LAD successfully grounds the generation in visual evidence, yielding descriptions that are both factually accurate and rich in detail.

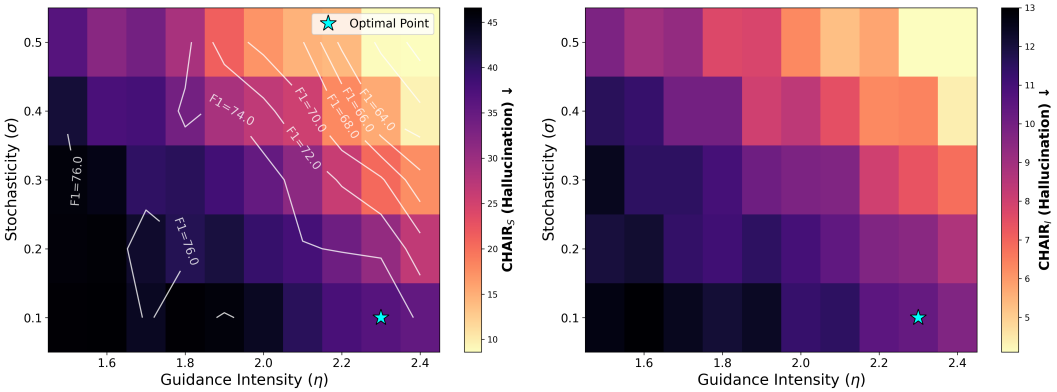


Figure 12: Hyperparameter sensitivity for Shikra. Sharing the dense linear projection architecture with LLaVA, Shikra shows a similar tolerance for high guidance strengths. The landscape indicates higher sensitivity to stochasticity, with peak F_1 scores clustered at lower σ. Our chosen point ($\eta = 2.3, \sigma = 0.1$) effectively minimizes CHAIR_S while adhering to the high-F_1 ridge.

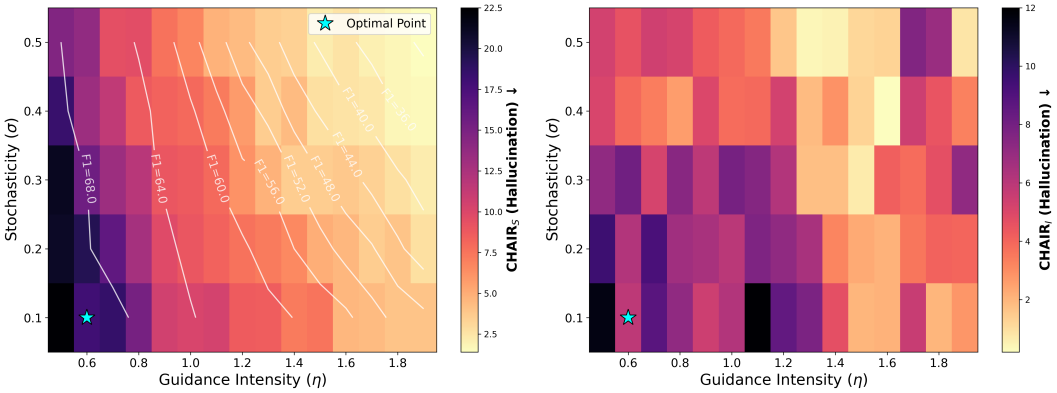


Figure 13: Hyperparameter sensitivity for MiniGPT-4. Due to the information bottleneck imposed by the Q-Former (compressing vision into $\sim$32 tokens), the landscape is notably sharper. Hallucination rates drop precipitously even at low guidance levels, and the tight F_1 contours indicate a narrow error margin. Our selection ($\eta = 0.6, \sigma = 0.1$) uses lower guidance to avoid over-correction.

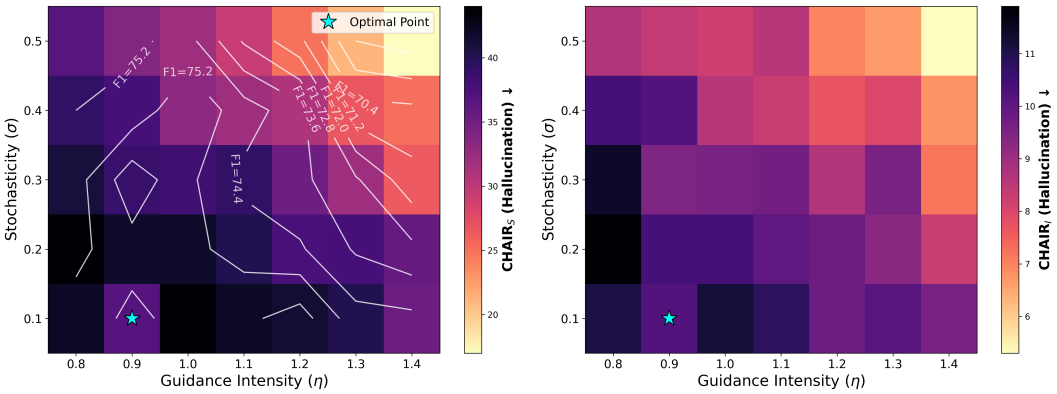


Figure 14: Hyperparameter sensitivity for InstructBLIP. Also utilizing a Q-Former, this profile exhibits a localized convex peak around $\eta \approx 0.9$, distinct from the linear ridges in dense projection models. Our configuration ($\eta = 0.9, \sigma = 0.1$) precisely targets this maximum to ensure optimal discriminative precision without linguistic degradation.

LLaVA1.5-7B

- The image depicts a group of four people walking together in a grassy field, enjoying a pleasant day outdoors. They are walking in a line, with one person leading the way and the others following closely behind. The group appears to be a family, as they are all smiling and enjoying their time together. The field is surrounded by trees, providing a serene and natural setting for their walk. The people are carrying backpacks, which suggests they might be on a hiking or camping trip. The overall atmosphere of the scene is one of happiness and togetherness, as the family spends quality time together in the outdoors.

LLaVA1.5-7B+LAD

- The image shows a group of people walking through a lush green field, enjoying the outdoors. There are three people in the group, with two men and a woman. They are walking together, enjoying the beautiful landscape. The field is filled with green grass, and the sky is blue and clear. The group is walking in the grassy field, and they are enjoying the outdoors.

(a) LAD corrects the hallucinated face of people and removes fabricated details like “family” and “backpacks”.

LLaVA1.5-7B

- The image features a young boy and a woman standing on a beach near a lake. The boy is wearing a yellow shirt and appears to be looking at something in the water. The woman is standing next to him, possibly observing the same object or simply enjoying the view. In the background, there are a few boats visible on the water, adding to the serene atmosphere of the scene. The beach is a popular spot for relaxation and leisure activities, and the presence of the boats suggests that it might be a popular location for boating enthusiasts.

LLaVA1.5-7B+LAD

- The image features a woman and a young boy standing on a beach near a lake. The woman is wearing a grey shirt, while the boy is wearing a yellow shirt. They are both looking at the water, possibly admiring the view or taking pictures. The woman is wearing red shoes, while the boy is wearing white shoes. They are both standing on the shore, enjoying the beautiful lake.

(b) LAD removes the hallucinated “boats” from the background, adhering strictly to the visual content.

LLaVA1.5-7B

- The image features a group of men playing soccer on a grassy field. There are three main players visible in the scene, with one player standing closer to the left side of the field, another in the middle, and the third player on the right side. They are all wearing soccer uniforms and are actively engaged in the game. In addition to the main players, there are two more people in the background, one near the center and the other on the right side of the field. A sports ball can be seen in the middle of the field, indicating that the game is in progress.

LLaVA1.5-7B+LAD

- The image shows a group of men standing on a soccer field. They are wearing soccer uniforms and are likely playing a game of soccer. The men are standing on the field, with one of them wearing a red and white uniform. The other men are wearing white and red uniforms. They are all standing on the field, possibly waiting for the game to start.

(c) LAD corrects the hallucination of a “sports ball” and “two more people”, accurately describes the players’ state as standing, not actively playing.

Figure 15: Qualitative examples for LLaVA-1.5-7B.

Shikra-7B

- The image captures an exciting soccer game, with two players on the field chasing a soccer ball. One player is wearing a red uniform, while the other is dressed in a white uniform, likely representing the opposing team. Both players are in close proximity, showcasing their competitive spirit. In addition to the two main players, there are two other individuals visible in the scene, one on the left side and the other on the right side of the image. These two people seem to be teammates or opponents keeping an eye on the soccer match.

Shikra-7B+LAD

- The image captures a lively contested soccer match between two soccer players, one in red and the other in white. The two men are in a race to reach the soccer ball, which is located in the center of the scene. Both players are focused on the sports ball, showing their competitive spirit during the match.

(a) LAD produces a more focused description, removing the vaguely hallucinated "two other individuals" mentioned by the baseline.

Shikra-7B

- The image depicts a young girl walking down a rural road, enjoying her day outdoors. She appears to be wearing a blue jacket and is holding a frisbee in her hand, indicating that she might be playing frisbee on the road. In the background, there are several cars parked, further down the road. The scene conveys a sense of peacefulness and outdoor recreation, showcasing the girl's engaging activity on this quiet road.

Shikra-7B+LAD

- The image shows a young girl running along a paved road, enjoying a peaceful day in the countryside. She appears to be walking or running down the road, with her arms outstretched as if she's happily skipping down the hill. In the background, there are several other people, possibly walking or running, further down the road. The scene is lively, with various elements contributing to the overall lively atmosphere of the image.

(b) LAD avoids fabricating objects like a "frisbee" and "cars", instead correctly describing the girl's action.

Shikra-7B

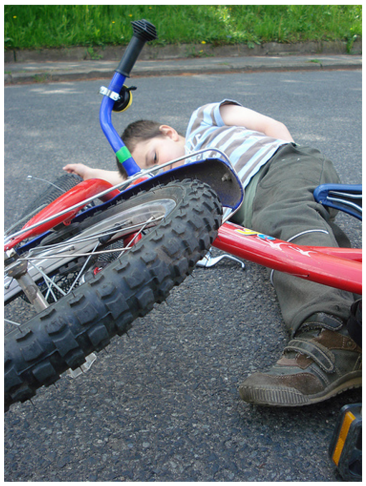
- The image captures a person riding a red motorcycle on a road. The rider is fully clothed and wearing a helmet for safety. The motorcycle is positioned in the middle of the scene, with both the rider and the vehicle occupying the majority of the frame. Although the rider is the primary subject of the image, there is another person visible near the right edge of the frame. Additionally, there are two potted plants placed along the top of the frame, adding a touch of greenery to the scene.

Shikra-7B+LAD

- The image captures a man riding a red motorcycle on a road. He is wearing an orange suit and helmet, indicating that he is likely participating in a motorcycle race. The rider is bending low to take a sharp turn on the red motorcycle. The motorcycle is one of the main focus of the scene.

(c) LAD correctly identifies the context (motorcycle race) and avoids hallucinating extraneous background details like "potted plants" and "another person".

Figure 16: Qualitative examples for Shikra-7B.



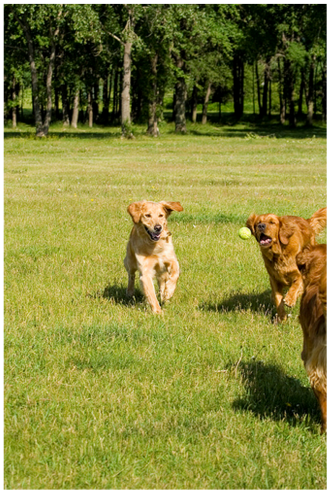
MiniGPT-4

- This image shows a boy lying on the ground next to a bicycle. The boy is wearing a striped shirt and has his head resting on his arm. The bicycle is lying on its side with its front wheel in the air. There is a road in the background with some trees and a building on the other side.

MiniGPT-4+LAD

- This image shows a boy lying on the ground next to a bicycle. The boy is wearing a striped shirt and has his head resting on the handlebars of the bike. The bike is lying on the ground with its wheels in the air.

(a) LAD provides a more focused description, removing the hallucinated background details of "trees and a building".



MiniGPT-4

- This image shows three dogs running through a green field. Two of the dogs are chasing a frisbee, while the third dog is running alongside them. The trees in the background are visible, and the sky is clear and blue.

MiniGPT-4+LAD

- This image shows three dogs running through a green field. Two of the dogs are chasing a ball, while the third is running alongside them. The trees in the background are visible through the grass.

(b) LAD corrects a classic object confusion by identifying a "ball" instead of a "frisbee", and avoids hallucinating details about the "sky".

MiniGPT-4

- This image shows a young boy sitting on a branch of a tree in a grassy field. The boy is wearing a yellow shirt and blue shorts and is holding onto the branch with his hands. The tree is a large, gnarled tree with branches that stretch out to the side. The grass in the field is tall and green, and there are some small flowers growing in the area around the tree. The sky is clear and blue, and there are some clouds in the distance.

MiniGPT-4+LAD

- This image shows a young boy sitting on a tree branch in a green field. The boy is wearing a yellow shirt and blue pants, and is holding onto the branch with his hands. The tree is tall and has a large branch that the boy is sitting on. The ground is covered in grass and there are some white flowers growing in the area around the tree.

(c) LAD improves attribute accuracy (e.g., "pants" not "shorts") while also avoiding the baseline's fabricated description of the "sky and clouds".

Figure 17: Qualitative examples for MiniGPT-4.

InstructBLIP

- The image features a single bed with a mattress sitting on a wooden floor. The mattress is unmade, and the bed frame is visible in the background. The bed is positioned in a corner of the room, with a wooden headboard and footboard visible. The room appears to be a bedroom, with a window in the background.

InstructBLIP+LAD

- A small, single bed is positioned in a room with a white mattress and a wooden headboard. The mattress is neatly placed on the floor, with its cover still on. There is a wooden headboard nearby, providing a comfortable sleeping surface for the bed. The room also features a wooden floor, which complements the overall aesthetic of the room.

(a) LAD accurately describes the room's interior, eliminating the hallucinated "window" in the background that the baseline inferred from context.

InstructBLIP

- The image features three birds perched on top of three tree stumps that are partially submerged in a body of water. The birds are standing on the stumps, which are partially exposed above the water level. The birds are of different sizes, with one being slightly larger than the other two. They are all facing in different directions, adding to the dynamic composition of the scene. The water is calm and serene, with a few small waves lapping against the shore. The overall atmosphere of the image is peaceful and natural, with the birds and the water creating a harmonious setting.

InstructBLIP+LAD

- The image features a group of three birds perched on top of three tree stumps, standing on the edge of a body of water. The birds are perched on the tree stumps, which are partially submerged in the water. They are positioned in a row, with one bird on the left, another in the middle, and the third on the right. The birds are all facing the same direction, possibly enjoying the view of the water or the surrounding area. The scene is set in a natural environment, with the birds and the tree stumps providing a unique and interesting composition.

(b) LAD corrects the behavioral description (birds facing the *same* direction) and removes non-existent environmental details like "small waves".

InstructBLIP

- The image features a bench situated on a wooden deck overlooking the ocean. The bench is placed near the edge of the deck, providing a scenic view of the ocean and the sunset. The bench is positioned in a way that allows for a clear view of the water and the sunset. There are several palm trees in the background, adding to the tropical atmosphere of the scene. The bench is the focal point of the image, inviting visitors to sit and enjoy the serene atmosphere.

InstructBLIP+LAD

- The image features a bench sitting on a wooden walkway overlooking the ocean and a palm tree. The bench is positioned near the ocean, providing a scenic view of the water and the palm tree. The sun is setting in the background, casting a warm and inviting glow over the area. The bench is empty, inviting visitors to sit and enjoy the peaceful atmosphere.

(c) LAD refines the setting by correctly identifying a "walkway" instead of a "deck" and accurately quantifying the single "palm tree", avoiding the baseline's pluralization hallucination.

Figure 18: Qualitative examples for InstructBLIP.