DYNAMIC STEERING OF LAYER-WISE UNCERTAINTY TO MITIGATE HALLUCINATION IN LVLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (LVLMs) suffer from diverse hallucinations, such as errors in object presence, spatial reasoning, counting, and text recognition. Existing inference-time mitigation methods largely focus on object hallucination and often apply coarse or uniform interventions, overlooking variations in hallucination risk across model depth. In this paper, we propose a dynamic steering method based on uncertainty that adaptively controls intervention strength at each decoder layer. Our analysis reveals that hallucination patterns vary substantially across both decoder layers and models, and no single attention-derived signal consistently captures hallucination risk. To overcome this variability, we integrate multiple complementary attention-derived signals to estimate layer-wise hallucination uncertainty as a dynamic control signal for steering. It applies stronger interventions to high-risk layers while preserving reliable representations, enabling adaptive steering across both decoder layers and generation steps without additional training. Experiments on multiple LVLMs across four hallucination benchmarks show that our method consistently improves performance across various hallucination tasks over existing inference-time baselines, achieving an average gain of 5%p and demonstrating that attention-derived layer-wise uncertainty provides an effective control signal for dynamically steering LVLMs against hallucinations beyond object-centric settings.

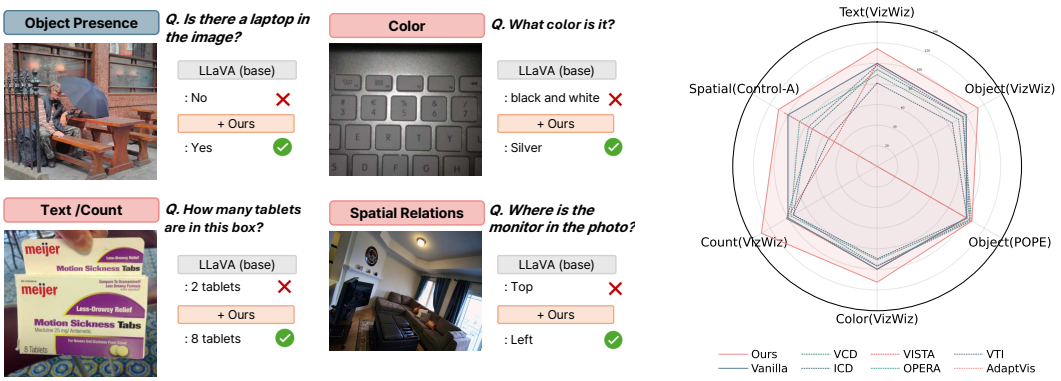
1 INTRODUCTION

Large vision–language models (LVLMs) have recently shown strong performance on multimodal generation and understanding tasks, such as visual question answering and image captioning. Yet, as illustrated in fig1(a), they still hallucinate often by describing nonexistent objects, misidentifying attributes (e.g., color or text), and producing incorrect counts or spatial relations.

Most existing hallucination mitigation methods such as contrastive decoding (Wang et al., 2024; Sennrich et al., 2024; Kim et al., 2025; Leng et al., 2024), steering-based interventions (Liu et al., 2025; Li et al., 2024), and reinforcement learning (Yu et al., 2026) are developed and evaluated primarily for object hallucinations, often guided by benchmarks such as POPE and CHAIR that focus on object existence and category errors (Yifan Li & Wen, 2023; Rohrbach et al., 2018a). As a result, hallucinations beyond object existence (e.g., attribute, text/count, and relational errors) remain comparatively underexplored. As shown in Fig. 1(b), methods that improve object-centric benchmarks can even degrade performance when non-object hallucinations are prominent. This suggests that many current interventions are implicitly tuned to object hallucinations and do not generalize evenly across hallucination types.

Prior work often explains LVLM hallucinations through a language-model-centric lens, attributing them to strong linguistic priors that override visual evidence (Sennrich et al., 2024; Kim et al., 2025), and addresses them by adjusting outputs at the decoding stage, for example, via contrastive decoding (Leng et al., 2024) or activation steering (Liu et al., 2025; Li et al., 2024). However, these methods apply intervention uniformly across layers without accounting for variations in hallucination risk across model depth and generation steps. We instead analyze the model's internal attention patterns and find that hallucination-related behaviors vary substantially across decoder layers and models. In particular, image-token attention, text-token attention, attention entropy, and attention concentration exhibit different relationships with hallucination across layers and models, and no single attention-derived signal consistently captures hallucination risk. This variability motivates

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) (b)

Figure 1: **DUST reduces hallucinations.** (a) Examples by hallucination type: object, color, text/count, and spatial. (b) Some baselines drop below the base model on non-object benchmarks.

combining multiple complementary attention-derived signals to estimate layer-wise hallucination uncertainty, which can then serve as a dynamic control signal for modulating intervention strength while avoiding unnecessary corrections to reliable representations.

In this paper, we propose DUST (Dynamic Uncertainty-aware Steering), an uncertainty-aware hallucination mitigation method grounded in a layer-wise analysis of the model's internal attention behavior. DUST extracts multiple complementary attention-derived signals, including image-token attention, text-token attention, attention entropy, and attention concentration, and integrates them to estimate hallucination uncertainty for each decoder layer. By jointly considering these heterogeneous signals and their interactions, DUST captures layer-specific hallucination risk that cannot be reliably characterized by any single signal. We then perform uncertainty-aware adaptive steering that modulates the strength of hidden-state correction according to the estimated uncertainty, applying stronger corrections to high-risk layers while preserving reliable representations. This enables adaptive intervention across both decoder layers and generation steps without additional training.

DUST improves performance not only on object-hallucination benchmarks, but also on evaluations that explicitly target non-object hallucinations, achieving state-of-the-art results across multiple LVLMs. The key contributions of this paper are summarized as follows.

- **Layer-wise analysis of attention-derived hallucination signals**: We analyze internal attention behavior across decoder layers and models and show that hallucination-related patterns vary substantially, with no single attention-derived signal consistently capturing hallucination risk.
- **Multi-signal layer-wise uncertainty estimation**: We integrate multiple complementary attention-derived signals and their interactions to estimate hallucination uncertainty for each decoder layer, providing a more robust characterization of layer-specific hallucination risk.
- **Uncertainty-adaptive layer-wise steering**: We use the estimated layer-wise uncertainty as a dynamic control signal to modulate steering strength, enabling stronger intervention in high-risk layers while preserving reliable representations and mitigating both object and non-object hallucinations.

2 Related Work

2.1 Hallucination in LVLMs

LVLMs extend LLMs to multimodal inputs. However, they may produce hallucinations when images are incomplete or noisy, often relying on language priors rather than visual evidence(Leng et al., 2024; Yifan Li & Wen, 2023). Hallucinations include both object-level errors and non-object errors such as misreading text or misinterpreting attributes and spatial relations(Sun et al., 2024; Huang et al., 2024b). While existing mitigation methods (e.g., RL-based optimization(Sun et al., 2024; Yu et al., 2023), steering(Liu et al., 2025; Wu et al., 2025; Yang et al., 2024), and decoding-time

strategies(Leng et al., 2024; Wang et al., 2024; Sennrich et al., 2024; Kim et al., 2025)) have reduced object hallucinations, they tend to degrade performance on non-object hallucination tasks, suggesting that object-centric optimization does not generalize across hallucination types. Addressing diverse hallucinations in LVLM therefore requires mitigation guided by the model's internal uncertainty rather than task-specific interventions.

2.2 Uncertainty Quantification

Uncertainty Quantification (UQ) measures how reliable a model's predictions are by estimating predictive uncertainty (Gawlikowski et al., 2021). In LLMs, concerns about reliability, including hallucinations, have motivated extensive UQ research. Existing methods quantify uncertainty using signals such as layer-wise consistency (e.g., projecting intermediate hidden states to the vocabulary/logit space)(Orgad et al., 2025; Chuang et al., 2023; Geva et al., 2020), attention- and representation-based measures (e.g., attention concentration or geometric properties)(Binkowski et al., 2025; Chuang et al., 2024; Marks & Tegmark, 2024), and response variability under repeated sampling (Farquhar et al., 2024; Manakul et al., 2023; Kuhn et al., 2023). In contrast, UQ for LVLMs remains relatively underexplored, and it is unclear whether these uncertainty signals are sufficient on their own or require integration to reliably estimate hallucination risk. This work examines how attention-based uncertainty signals can be combined to enable internal uncertainty estimation without an external classifier.

3 Preliminary Analysis

In this section, we present preliminary experiments to analyze hallucination in LVLMs, with a focus on the behavior of the LLM decoder.

3.1 Layer-wise Probing of Uncertainty Signals

To identify where hallucination-related signals emerge, we probe diverse uncertainty metrics across decoder layers, including logits, hidden states, attention dynamics, layer-mismatch indicators, and multi-sample uncertainty statistics. For each metric, we compute layer-wise AUROC to measure its ability to distinguish hallucinated outputs. Figure 2 shows that hallucination signals are highly metric-dependent and vary across depths. Most metrics provide only weak or localized separation, whereas attention-head entropy exhibits a clear and stable signal, especially in the middle-to-late decoder layers. This suggests that hallucination is more strongly reflected in head-wise attention uncertainty than in token-level confidence or aggregate representation statistics. This trend is consistent across model families. As shown in the Appendix, LLaVA, InstructBLIP, and Qwen all show that attention-head-wise entropy produces the most prominent and stable AUROC patterns, while other metrics such as logit entropy, top-k mass, hidden-state statistics, and head-similarity measures are less consistent. Overall, these results identify attention-head entropy as the most informative single signal among the candidates, yet its separability remains moderate and depends on layer depth, suggesting that entropy alone is insufficient to characterize hallucination risk.

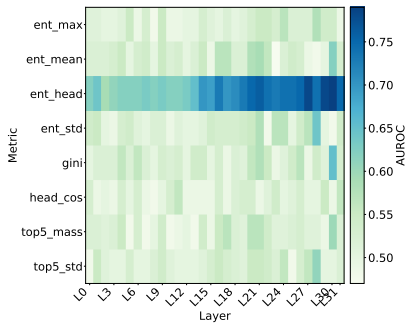


Figure 2: Layer-wise AUROC of candidate uncertainty metrics for hallucination detection.

3.2 Head-wise Pattern Analysis

To analyze hallucination behavior at the attention-head level, we characterize each head using changes in attention entropy and attention mass on image and question tokens. For each layer ℓ and head h, we compute

$$\begin{aligned}\Delta E_{\ell,h} &= \bar{H}^{\text{hall}}_{\ell,h} - \bar{H}^{\text{corr}}_{\ell,h},\\ \Delta M^{\text{img}}_{\ell,h} &= \bar{M}^{\text{img,hall}}_{\ell,h} - \bar{M}^{\text{img,corr}}_{\ell,h},\\ \Delta M^{\text{q}}_{\ell,h} &= \bar{M}^{\text{q,hall}}_{\ell,h} - \bar{M}^{\text{q,corr}}_{\ell,h}.\end{aligned} \tag{1}$$

Based on these changes, each head is categorized into one of five attention patterns: target shift without entropy change (purple), diffuse grounding failure (red), diffuse but grounded (green), focused wrong anchor (orange), or no clear change (blue). Red heads correspond to increased entropy with reduced image-token attention, green heads to increased entropy while preserving image-token attention, and orange heads to more concentrated attention with reduced question-token attention. Figure 3 reveals distinct head-wise patterns across models. LLaVA exhibits many diffuse-but-grounded heads in the middle layers, whereas InstructBLIP shows more diffuse grounding failures. These results indicate that hallucination is not characterized by attention entropy alone, but by joint changes in attention diffuseness and grounding-related attention allocation, motivating the use of multiple complementary attention signals for uncertainty estimation.

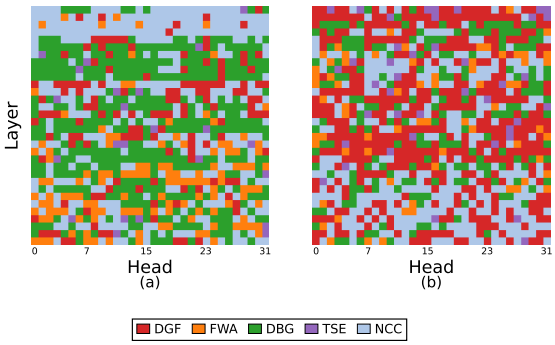


Figure 3: Head-wise hallucination patterns for LLaVA and InstructBLIP.

4 Methodology

4.1 Problem Formulation

We consider a large vision-language model composed of a vision encoder and a language decoder. Given an image x, the vision encoder extracts visual features $f = E_v(x)$. Conditioned on the visual features f, a text prompt q, and previously generated tokens $y_{<t}$, the decoder predicts the next-token distribution

$$p_\theta(y_t \mid y_{<t}, q, f). \tag{2}$$

Hallucination occurs when the generated response contains content that is not supported by the visual input. We formulate hallucination analysis as an internal uncertainty estimation problem. Rather than relying on a single signal such as entropy or confidence, we characterize uncertainty using multiple attention patterns that capture attention diffuseness, visual grounding, textual reliance, and concentration.

At decoding step t and decoder layer l, let $\phi_{l,t}$ denote these attention-derived signals. We define the layer-wise uncertainty score as

$$R_{l,t} = g(\phi_{l,t}), \tag{3}$$

where $R_{l,t} \in [0, 1]$ represents the estimated hallucination uncertainty at layer l. The function $g(\cdot)$ integrates heterogeneous and potentially interacting attention cues into a unified uncertainty measure. We use $R_{l,t}$ to adaptively control the intervention strength, applying stronger corrections to high-uncertainty states while minimizing modifications to reliable representations.

4.2 Dynamic Uncertainty-aware Steering

Figure 4 presents an overview of our method, which consists of three stages: (1) visual direction extraction, (2) layer-wise uncertainty estimation, and (3) dynamic steering.

4.2.1 Visual Direction Extraction

We first extract a layer-wise visual steering direction that captures how decoder representations are shifted by visual conditioning. For each calibration sample, we construct a paired input by processing

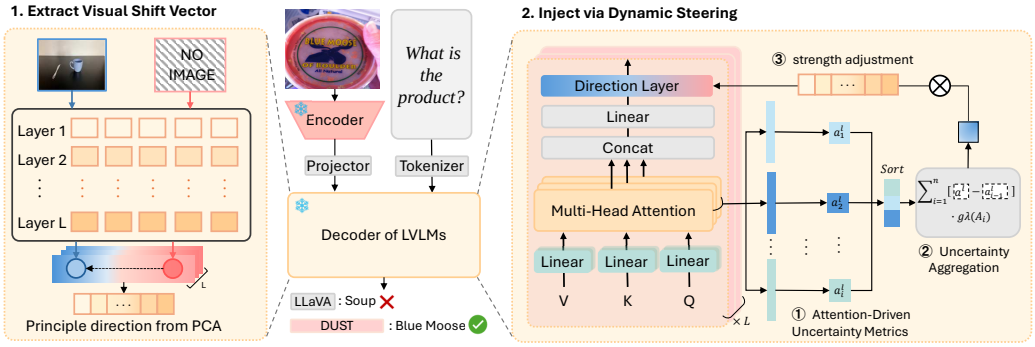


Figure 4: **Overview of the proposed Dynamic Steering method.** It illustrates the three-step process: (i) visual direction extraction via PCA, (ii) layer-wise uncertainty estimation by integrating multiple attention-derived signals, and (iii) uncertainty-aware steering modulation within the LVLM decoder.

the same text prompt q with and without the original image I. Let $h^{(l)}(I, q)$ and $h^{(l)}(\tilde{I}, q)$ denote the decoder hidden states at the last token position of layer l under the image-conditioned and no-image conditions, respectively, where $\tilde{I}$ denotes the absence of visual input. We define the visual-induced representation shift as

$$d^{(l)} = h^{(l)}(I, q) - h^{(l)}(\tilde{I}, q). \tag{4}$$

For each decoder layer, we collect these shifts over calibration samples and estimate the dominant visual-conditioning direction by applying PCA. Let $\mathbf{u}_{l,1}$ denote the first principal component of the collected shifts at layer l. Since the sign of a principal component is inherently ambiguous, we orient $\mathbf{u}_{l,1}$ to have positive alignment with the mean representation shift:

$$\mathbf{u}_{l,1}^{\top} \bar{d}^{(l)} \geq 0, \tag{5}$$

where $\bar{d}^{(l)}$ denotes the mean shift over the calibration samples. We then normalize the oriented component to obtain the layer-wise visual steering direction:

$$v_l = \frac{\mathbf{u}_{l,1}}{\|\mathbf{u}_{l,1}\|_2}. \tag{6}$$

The visual directions are estimated from 50 randomly sampled benchmark instances without using labels or hallucination annotations, using only internal activations. This direction represents the dominant change in decoder representations induced by the presence of visual input. During inference, it serves as the layer-specific direction along which the model is steered, while the steering strength is adaptively controlled by the hallucination uncertainty score.

4.2.2 Layer-wise Uncertainty Estimation

As shown in Section 3.2, hallucination is associated with joint changes in attention diffuseness and grounding-related attention allocation. We therefore characterize each decoder layer using four complementary attention cues:

- $H_t^{(l)}$: mean attention entropy across heads, measuring attention diffuseness.
- $M_t^{(l),\text{img}}$: mean attention mass on image tokens, measuring visual attention allocation.
- $M_t^{(l),\text{q}}$: mean attention mass on question tokens, measuring reliance on textual context.
- $C_t^{(l)}$: mean per-head maximum attention probability, measuring attention concentration.

Because these quantities have different scales across layers, we standardize each cue using layer-wise calibration statistics:

$$z_{x,t}^{(l)} = \frac{x_t^{(l)} - \mu_x^{(l)}}{\sigma_x^{(l)}}. \tag{7}$$

We then map the standardized cues to fuzzy membership values in $[0, 1]$:

$$E^+_{l,t} = \sigma\left(z^{(l)}_{H,t}\right), \qquad I^-_{l,t} = \sigma\left(-z^{(l)}_{M^{\text{img}},t}\right), \tag{8}$$

$$Q^+_{l,t} = \sigma\left(z^{(l)}_{M^{\text{q}},t}\right), \qquad C^+_{l,t} = \sigma\left(z^{(l)}_{C,t}\right). \tag{9}$$

Here, E^+, I^-, Q^+, and C^+ represent diffuse attention, weak visual attention, increased attention to the question context, and concentrated attention, respectively. Their complementary states are defined as $E^- = 1 - E^+$ and $I^+ = 1 - I^-$.

To capture interactions among these cues, we combine their memberships into four attention-state patterns:

$$\begin{aligned} S^{\text{DL}}_{l,t} &= \min(E^+_{l,t}, I^-_{l,t}), \quad S^{\text{DH}}_{l,t} = \min(E^+_{l,t}, I^+_{l,t}), \\ S^{\text{SF}}_{l,t} &= \min(E^-_{l,t}, C^+_{l,t}), \quad S^{\text{TD}}_{l,t} = \min(Q^+_{l,t}, I^-_{l,t}), \end{aligned} \tag{10}$$

where DL, DH, SF, and TD denote diffuse-low-image, diffuse-high-image, sharp-focus, and text-dominant attention states, respectively. Rather than treating any single pattern as a direct indicator of hallucination, we use their joint configuration to estimate layer-wise hallucination uncertainty. We define the pattern vector as

$$\mathbf{s}_{l,t} = \left[S^{\text{DL}}_{l,t}, S^{\text{DH}}_{l,t}, S^{\text{SF}}_{l,t}, S^{\text{TD}}_{l,t}\right], \tag{11}$$

and compute the uncertainty score as

$$R_{l,t} = \mathcal{A}\left(\mathbf{s}_{l,t}\right), \tag{12}$$

where $\mathcal{A}$ aggregates the four attention-state patterns into a single score. We use a symmetric λ-Choquet integral to account for interactions among these patterns.

4.2.3 Dynamic Steering

At inference time, the uncertainty score $R_{l,t}$ serves as a layer-wise steering gate. Let $m^{(l)}_t$ denote the MLP output at layer l and decoding step t, with unit direction

$$\hat{m}^{(l)}_t = \frac{m^{(l)}_t}{\|m^{(l)}_t\|_2}. \tag{13}$$

We adjust its direction according to the estimated uncertainty:

$$\tilde{m}^{(l)}_t = \hat{m}^{(l)}_t + R_{l,t} v_l, \tag{14}$$

and preserve the original activation magnitude by normalization:

$$m'^{(l)}_t = \|m^{(l)}_t\|_2 \frac{\tilde{m}^{(l)}_t}{\|\tilde{m}^{(l)}_t\|_2}. \tag{15}$$

Thus, higher uncertainty induces stronger steering toward the visual direction, while lower uncertainty preserves the original activation direction. Since $R_{l,t}$ is estimated for each layer and decoding step, the steering dynamically adapts across layers and generation steps without additional training.

5 Experiments

5.1 Experimental Setting

Experimental Setup. We evaluate Dynamic Steering on multiple LVLMs, including LLaVA-1.5 (Liu et al., 2024) and InstructBLIP (Dai et al., 2023). DUST operates in a training-free manner: visual steering directions are estimated from calibration samples and fixed throughout inference, while the steering strength is dynamically adjusted based on attention-derived uncertainty. Specifically, we derive four internal cues from attention patterns(attention entropy, image-token attention mass, question-token attention mass, and attention concentration) and aggregate them using a λ-Choquet integral to construct a layer-wise uncertainty gate.

Table 1: **Overall evaluation of DUST.** Left: generalization across different LVLM backbones. Right: accuracy-based comparison on object and non-object hallucination benchmarks.

(a) Generalizations

Model	Acc.	Δ
LLaVA1.5-7B	82.90	–
DUST	**88.07**	+5.17
LLaVA1.6-7B	89.17	–
DUST	**89.90**	+0.73
InstructBLIP	80.63	–
DUST	**86.80**	+5.97
Qwen2-VL-7B	89.83	–
DUST	**92.37**	+2.54

(b) Accuracy comparison across hallucination benchmarks

Method	Object		Non-Object				Avg	Δ
	POPE	VizWiz	Spatial	Text	Color	Count		
Vanilla	80.83	39.90	65.66	33.76	41.43	29.01	48.43	–
VCD	84.66	38.28	56.22	31.75	40.11	29.43	46.74	-1.69
ICD	80.06	33.62	39.55	27.34	37.16	27.34	40.84	-7.59
VISTA	84.30	39.68	30.03	33.80	41.31	29.43	43.09	-5.34
OPERA	84.82	36.39	64.73	30.08	37.69	28.03	46.95	-1.48
VTI	85.02	38.44	55.39	33.10	40.07	27.20	46.53	-1.90
DUST	**86.08**	**45.14**	**68.19**	**38.54**	**46.62**	**37.66**	**53.70**	+5.27

Table 2: **Comparison on the POPE benchmark across LVLMs.**

Method	LLaVA-1.5			InstructBLIP		
	Random	Popular	Adv.	Random	Popular	Adv.
Vanilla	82.90	81.00	78.60	80.63	80.50	77.40
VCD	87.73	85.38	80.88	80.63	81.47	79.56
ICD	83.66	80.46	76.07	86.43	82.93	80.87
VISTA	83.52	85.87	83.52	86.01	84.43	82.20
OPERA	85.97	85.26	83.23	**89.03**	80.93	76.33
VTI	86.30	85.27	83.50	86.60	84.90	82.80
DUST	**88.07**	**86.47**	**83.70**	86.80	**85.40**	**82.92**

Table 3: **Results across object- and attribute-level hallucinations.** Scores are reported for existence, count, position, and color.

Method	Object-Level		Attribute-Level		Total Scores
	Existence	Count	Position	Color	
Vanilla	171.67	106.67	80.00	160.00	518.34
VCD	184.66	138.33	128.67	153.00	604.66
ICD	185.00	131.66	113.33	138.33	568.32
VISTA	85.00	96.66	58.33	65.00	304.99
OPERA	195.00	153.33	133.33	155.00	636.66
VTI	190.00	120.00	138.33	165.00	613.33
DUST	190.00	163.33	133.33	158.33	**644.99**

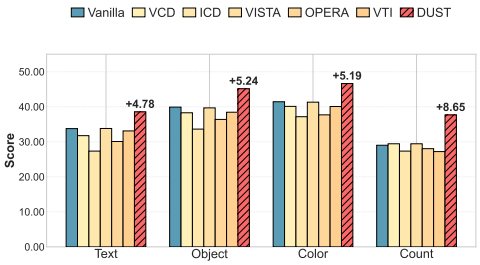


Figure 5: **Comparison of visual skill performance on the VizWiz dataset across four visual-skill categories.** DUST consistently improves scores over the Vanilla baseline.

Table 4: **Ablation study on fusion strategies.** Different fusion strategies are compared for aggregating attention-state patterns into the uncertainty score.

Fusion	POPE-R	POPE-P	POPE-A	VizWiz	Avg.
Vanilla	82.90	81.00	78.60	38.14	70.16
Max	87.56	86.13	83.49	38.54	73.93
Min	87.08	85.89	83.47	43.25	74.92
Median	87.38	85.99	83.58	41.11	74.52
Power Mean	87.32	85.91	83.49	40.42	74.29
Choquet	**88.07**	**86.47**	**83.70**	**43.55**	**75.45**

We compare DUST against representative inference-time hallucination mitigation methods, including contrastive decoding approaches such as VCD (Leng et al., 2024), ICD (Wang et al., 2024), and OPERA (Huang et al., 2024a), as well as steering-based approaches such as VISTA (Li et al., 2024) and VTI (Liu et al., 2025). For evaluation, we consider both object-centric and non-object hallucinations. We use POPE (Yifan Li & Wen, 2023) and MME (Fu et al., 2025) to evaluate object-centric hallucinations, while leveraging What's Up (Kamath et al., 2023) and VizWiz (Gurari et al., 2018) to evaluate non-object hallucinations involving broader visual understanding.

5.2 Experimental Results

Main Results. Table 1 summarizes the generalization of DUST across LVLM backbones and hallucination categories. Across four LVLMs, DUST consistently improves the vanilla baseline, with gains ranging from $+0.73$ to $+7.00$ points, demonstrating that its effectiveness is not specific to a particular architecture. Across six object and non-object hallucination categories, DUST also improves over vanilla in every category, increasing the average score from 48.43 to 53.70 ($+5.27$). In contrast, all competing inference-time mitigation methods obtain lower overall averages than vanilla,

(a) AMBER

Method	AMBER			
	Acc.	Precision	Recall	F1
Vanilla	81.20	83.90	88.70	86.20
VCD	77.10	82.90	82.50	82.70
ICD	76.80	85.60	78.10	81.80
VISTA	80.60	83.50	88.20	85.80
OPERA	75.50	94.10	67.60	78.70
VTI	78.70	80.00	90.60	85.00
DUST	**82.30**	88.00	84.90	**86.40**

(b) What's Up

Method	C-A	C-B	COCO	VG	Avg.
Vanilla	61.08	61.03	62.52	39.27	56.99
VCD	33.63	31.87	53.81	39.61	38.34
ICD	20.68	12.09	42.22	31.54	24.58
VISTA	20.39	16.67	27.75	12.21	19.11
OPERA	40.13	31.75	**64.73**	**52.28**	44.97
VTI	31.72	42.81	57.55	30.95	40.06
DUST	**78.32**	**66.83**	62.79	41.18	**64.34**

Figure 6: **Evaluation on AMBER and spatial reasoning benchmarks.** (a) Performance on the AMBER discriminative task. (b) Performance across spatial reasoning benchmarks.

indicating that improvements in their target hallucination type do not necessarily transfer across categories.

DUST remains effective on conventional object-centric hallucination benchmarks. On POPE (Table 2), DUST improves over vanilla under Random, Popular, and Adversarial settings for both LLaVA-1.5 and InstructBLIP, achieving the best result in five of the six settings. On the object- and attribute-level evaluation (Table 3), DUST achieves the highest total score of 644.99. Although some competing methods perform better on individual sub-tasks, DUST maintains strong performance across existence, count, position, and color rather than relying on gains from a single category. Similarly, on AMBER (Figure 6(a)), DUST achieves the highest accuracy (82.30) and F1 score (86.40), showing balanced object-level discrimination despite other methods performing better on individual precision or recall metrics.

More importantly, the benefit of DUST extends beyond object-centric hallucination. On VizWiz (Figure 5), DUST improves over vanilla on Text, Object, Color, and Count by $+4.78$, $+5.24$, $+5.19$, and $+8.65$ points, respectively, whereas existing mitigation methods frequently degrade performance in these categories. A similar trend appears in spatial reasoning (Figure 6(b)): DUST improves over vanilla across all four controlled and real-world settings, increasing the average score from 56.99 to 64.34, while all competing methods obtain lower averages than vanilla. Thus, although DUST is not the best method on every individual sub-task, it avoids the systematic cross-category degradation observed in existing object-centric mitigation methods. Overall, these results show that DUST consistently mitigates both object and non-object hallucinations, rather than improving one hallucination type at the expense of others.

Fusion Strategy. Table 4 compares different strategies for aggregating attention-derived uncertainty cues. Conventional fusion methods, including Max, Min, Median, and Power Mean, improve over vanilla on average but show varying performance across benchmarks. In contrast, the Choquet integral consistently performs best across the evaluated POPE settings and VizWiz, achieving the highest average score of 75.45. This result supports jointly modeling interactions among attention-derived uncertainty cues, rather than treating each cue independently.

Ablation and Efficiency Analysis. Table 4 evaluates different strategies for aggregating attention-derived uncertainty cues. Although conventional fusion rules improve over vanilla on average, their performance varies across benchmarks. The Choquet integral achieves the highest average score of 75.45, supporting the use of interaction-aware fusion rather than treating the uncertainty cues independently.

Figure 7(a) further analyzes the proposed uncertainty-aware steering components. Fixed PCA substantially improves POPE but decreases VizWiz from 38.14 to 36.67, showing that a static steering direction can improve object hallucination while degrading performance on other visual capabilities. In contrast, uncertainty-aware variants improve both POPE and VizWiz overall. Using individual cues already provides strong gains, while removing any cue from the full model reduces the average score from 75.45 to 74.19–74.75. Moreover, simple uniform averaging achieves 74.49, below the full model, indicating that both multiple uncertainty cues and their adaptive combination contribute to more consistent mitigation across benchmarks.

(a) Ablation Study

Method	POPE-R	POPE-P	POPE-A	VizWiz	Avg.
Vanilla	82.90	81.00	78.60	38.14	70.16
Fixed PCA	87.93	86.27	83.48	36.67	73.59
Entropy Only	87.34	86.07	83.57	**42.70**	74.92
Img. Attn. Only	**88.40**	**86.67**	83.70	40.51	74.82
Uniform Avg.	87.27	85.96	83.51	41.23	74.49
w/o Entropy	87.29	85.94	83.51	40.09	74.21
w/o Img. Attn.	87.43	86.09	83.62	41.43	74.64
w/o Q. Attn.	87.19	85.84	83.46	40.25	74.19
w/o Conc.	87.54	86.20	**83.71**	41.56	74.75
Full	88.07	86.47	83.70	43.55	**75.45**

(b) Accuracy-efficiency trade-off

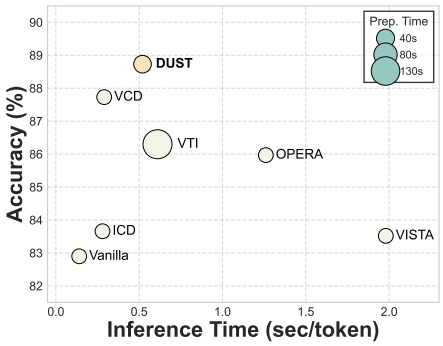


Figure 7: **Ablation and efficiency analyses.** (a) Ablation study of uncertainty-aware steering and attention cues. (b) Accuracy-efficiency trade-off across inference-time hallucination mitigation methods.

Figure 7(b) compares the accuracy–efficiency trade-off with existing inference-time mitigation methods. DUST lies in the high-accuracy, low-latency region while requiring little preparation overhead. In comparison, OPERA and VISTA incur substantially higher inference latency, while VTI requires larger preparation cost without achieving comparable accuracy. These results show that DUST improves hallucination mitigation without relying on expensive preprocessing or substantial additional inference cost.

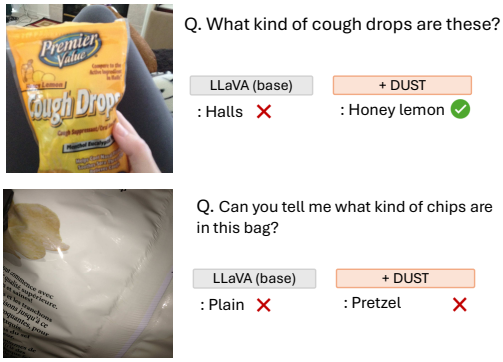


Figure 8: **Qualitative examples.** Comparison between the base model and DUST.

Qualitative Results. Figure 8 illustrates both the effectiveness and limitation of DUST. In the first example, the package contains visible evidence for the "Honey Lemon" flavor; the vanilla model hallucinates "Halls," whereas DUST correctly identifies the flavor. In the second example, the information required to determine the chip type is not sufficiently visible, and both models produce incorrect predictions. These examples suggest that DUST can better exploit available visual evidence to suppress hallucination, but cannot recover information that is absent or severely occluded in the input.

6 Concluding Remarks

This paper investigates the limitations of existing inference-time hallucination mitigation methods that mainly target object-centric errors and may degrade performance on other types of hallucinations. Our analysis shows that hallucination-related attention patterns vary across layers and models, making any single attention-derived signal insufficient for reliable uncertainty estimation. We therefore combine multiple complementary cues using the λ-Choquet integral and use the resulting layer-wise uncertainty to adaptively control visual steering. Experiments across multiple LVLMs and various hallucination benchmarks show consistent improvements on both object- and non-object hallucinations, without additional training and with low inference overhead. These results support attention-derived layer-wise uncertainty as an effective control signal to mitigate hallucinations beyond object-centric settings.

Limitations and Future Work. The proposed method requires access to intermediate attention maps and hidden states, limiting its use with closed-source or API-only LVLMs. Future work will investigate uncertainty estimation and steering strategies across broader LVLM architectures and multimodal tasks.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish the writing, including improving grammar, clarity, and readability. We did not use generative AI tools to generate research ideas, conduct analyses, interpret results, or produce research findings. Other disclosure tasks are not applicable to this work. All AI-assisted text was reviewed and revised by the authors to ensure accuracy and consistency with the intended meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Jakub Binkowski, Denis Janiak, Albert Sawczyn, Bogdan Gabrys, and Tomasz Jan Kajdanowicz. Hallucination detection in LLMs using spectral features of attention maps. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 24354–24385, 2025. URL https://aclanthology.org/2025.emnlp-main.1239/.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. doi: 10.48550/arXiv.2309.03883.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. URL https://aclanthology.org/2024.emnlp-main.84/.

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

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625 – 630, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*, 2025. URL https://openreview.net/forum?id=DgH9YCsqWm.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseo Lee, Matthias Humt, Jianxiang Feng, Anna M. Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, M. Shahzad, Wen Yang, Richard Bamler, and Xiaoxiang Zhu. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56:1513–1589, 2021.

Mor Geva, R. Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. URL https://doi.org/10.18653/v1/2021.emnlp-main.446.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 3608–3617, 2018.

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

Wen Huang, Hongbin Liu, Minxin Guo, and N. Gong. Visual hallucinations of multi-modal large language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024b. doi: 10.48550/arXiv.2402.14683.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9161–9175, 2023.

Jieun Kim, Jinmyeong Kim, Yoonji Kim, and Sung-Bae Cho. Fuzzy contrastive decoding to alleviate object hallucination in large vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20572–20581, 2025.

Lorenz Kuhn, Y. Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. URL https://doi.org/10.48550/arXiv.2302.09664.

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

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll ár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in neural information processing systems (NeurIPS)*, volume 36, 2024.

Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9004–9017, 2023. URL https://aclanthology.org/2023.emnlp-main.557.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *Proceedings of the First Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=KRnsX5Em3W.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018a. URL https://api.semanticscholar.org/CorpusID:52176506.

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

Rico Sennrich, Jannis Vamvas, and Alireza Mohammadshahi. Mitigating hallucinations and off-target machine translation with source-contrastive and language-contrastive decoding. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 21–33, 2024. URL https://aclanthology.org/2024.eacl-short.4.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics (ACL)*, pp. 13088–13110, 2024. URL https://aclanthology.org/2024.findings-acl.775.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

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

Jun Wu, Yue Ding, Guofan Liu, Tianze Xia, Ziyue Huang, Dianbo Sui, Qiang Liu, Shu Wu, Liang Wang, and Tieniu Tan. Sharp: Steering hallucination in lvlms via representation engineering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025. URL https://aclanthology.org/2025.emnlp-main.725/.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. pp. 14635–14645, 2024.

Kun Zhou Jinpeng Wang Wayne Xin Zhao Yifan Li, Yifan Du and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL https://openreview.net/forum?id=xozJw0kZXF.

Chengzhi Yu, Yifan Xu, Yifan Chen, and Wenyi Zhang. Optimizing lvlms with on-policy data for effective hallucination mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4733–4743, March 2026.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi: 10.1109/CVPR52733.2024.01310.

APPENDIX CONTENTS

A Experimental Setup 13

A.1 Implementation Details

A.2 Baselines

A.3 Benchmarks and Metrics

B λ-Choquet Aggregation 14

C Additional Results 15

C.1 Full POPE Results

C.2 CHAIR Open-ended Captioning Results

C.3 Full Spatial Reasoning Results

D Layer-wise Probe AUROC 16

A EXPERIMENTAL SETUP

A.1 IMPLEMENTATION DETAILS

We apply DUST to LLaVA-1.5-7B Liu et al. (2024), LLaVA-1.6-Mistral-7B, InstructBLIP with a Vicuna-7B language backbone Dai et al. (2023), and Qwen2-VL-7B-Instruct. Unless otherwise stated, LLaVA-1.5-7B is used for the analyses and ablations. DUST keeps the backbone parameters frozen; visual directions and cue statistics are estimated before inference, and the gate is recomputed from attention during inference. The implementation loads `liuhaotian/llava-v1.5-7b`, `llava-hf/llava-v1.6-mistral-7b-hf`, `Salesforce/instructblip-vicuna-7b`, and `Qwen/Qwen2-VL-7B-Instruct`, respectively. All experiments are conducted on NVIDIA A100 GPUs.

A.2 BASELINES

Contrastive decoding.

- **VCD** Leng et al. (2024): Contrasts output distributions from the original image and a visually distorted image to reduce language-prior-driven hallucinations.
- **ICD** Wang et al. (2024): Contrasts standard and disturbed instruction conditions to suppress hallucinated concepts.
- **OPERA** Huang et al. (2024a): A beam-search-based decoding method that penalizes over-reliance on summary tokens and applies a retrospection-allocation strategy.
- **FuzzyCD** Kim et al. (2025): Uses a Takagi–Sugeno fuzzy gate to adapt visual contrastive decoding between the original and perturbed image conditions.

Steering-based methods.

- **VISTA** Li et al. (2024): Steers internal activations to reinforce visual evidence during decoding.
- **VTI** Liu et al. (2025): Adjusts latent representations to improve visual-text alignment.

A.3 BENCHMARKS AND METRICS

POPE Yifan Li & Wen (2023). A binary yes/no benchmark for object existence hallucination under random, popular, and adversarial negative sampling. We use the MSCOCO image split, with 3,000 questions per sampling setting and equal numbers of positive and negative questions. We report accuracy, precision, and recall, with "yes" as the positive class. For counts TP, TN, FP, FN, accuracy is $(TP + TN)/N$, precision is $TP/(TP + FP)$, and recall is $TP/(TP + FN)$.

AMBER Wang et al. (2023). The full discriminative split contains 14,216 questions over 1,004 images (query IDs 1005–15220). We report accuracy, precision, recall, and $F_1 = 2PR/(P + R)$, with the positive-label convention specified by the evaluator.

MME Fu et al. (2025). We use the existence, count, position, and color subsets and report the official score for each subset and their sum. Each subset score is $100(\mathrm{Acc} + \mathrm{Acc}^{+})$, where Acc^{+} is the fraction of images for which both associated questions are answered correctly. Each subset has a maximum score of 200, and the four-subset total has a maximum of 800.

What's Up Kamath et al. (2023). A spatial reasoning benchmark in which images differ in object relations while preserving object identities. For Controlled-A and Controlled-B, we report per-image accuracy (Acc), pair accuracy (all images in a pair correct), and set accuracy (all images in a set correct). COCO and VG are natural-image subsets split into questions involving one or two objects. The condensed scores in Figure 6(b) average the metrics within each subset. The local question files contain 412, 408, 2,247, 440, 1,160, and 291 questions for Controlled-A, Controlled-B, COCO-one, COCO-two, VG-one, and VG-two, respectively. Controlled-A tests left/right and on/under; Controlled-B tests left/right and front/behind. Natural-image relations include horizontal, vertical, and, for VG, depth relations. Pair and set scores require all members of their annotated group to be correct; they cannot be recovered by averaging per-image scores.

VizWiz Gurari et al. (2018). A real-world VQA benchmark collected from blind or visually impaired users. We use the validation skill subset and its existing TXT (Text), OBJ (Object), COL (Color), and CNT (Count) tags. The local annotation file contains 2,245 unique image–question pairs and 5,830 category-tagged rows, including OTH. The category counts are 1,580 (Text), 2,244 (Object), 1,638 (Color), 239 (Count), and 129 (Other); categories overlap. The local scorer assigns $\min(m/3, 1)$ to a prediction matching m reference answers, and averages this score within each category. Thus, these values are consensus-based answer scores rather than binary exact-match accuracy.

CHAIR Rohrbach et al. (2018b). An open-ended captioning metric computed on 500 images sampled from MSCOCO Lin et al. (2014) val2014. The captioning runner samples without replacement using seed 42. The metrics are

$$\mathrm{CHAIR}_S = \frac{|\{\text{captions with} \geq 1 \text{ hallucinated object}\}|}{|\{\text{all captions}\}|}, \quad \mathrm{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}. \tag{16}$$

Recall is computed as the number of distinct ground-truth object categories mentioned per image, summed across images, divided by the summed number of ground-truth categories. Len. is the average number of NLTK-tokenized caption items before multiword object merging; it is not the language model's subword-token count. CHAIR and recall values in Table 6 are percentages. Ground-truth object categories are obtained from MSCOCO instance annotations and reference captions.

B λ-Choquet Aggregation

This section details the aggregation operator $\mathcal{A}$ in Eq. (12), which maps the pattern vector

$$s_{l,t} = [S_{l,t}^{\mathrm{DL}}, S_{l,t}^{\mathrm{DH}}, S_{l,t}^{\mathrm{SF}}, S_{l,t}^{\mathrm{TD}}] \in [0, 1]^4 \tag{17}$$

to the uncertainty score $R_{l,t}$.

Choquet integral. Let π be a permutation such that $s_{\pi(1)} \leq \cdots \leq s_{\pi(n)}$ with $n = 4$ (ties broken arbitrarily), and let $A_{(i)} = \{\pi(i), \ldots, \pi(n)\}$. Then

$$R_{l,t} = \sum_{i=1}^{n} \left[s_{\pi(i)} - s_{\pi(i-1)}\right] \mu_\lambda(A_{(i)}), \qquad s_{\pi(0)} = 0, \tag{18}$$

where μ_λ is a fuzzy measure on $\Omega = \{\mathrm{DL}, \mathrm{DH}, \mathrm{SF}, \mathrm{TD}\}$ with $\mu_\lambda(\emptyset) = 0$ and $\mu_\lambda(\Omega) = 1$. Since $s_{l,t} \in [0, 1]^4$, $R_{l,t} \in [0, 1]$.

Symmetric Sugeno λ-measure. We assign the same density to every pattern, so $\mu_\lambda(A)$ depends only on $|A|$. For $\lambda \neq 0$,

$$g_\lambda = \frac{(1 + \lambda)^{1/n} - 1}{\lambda}, \qquad \mu_\lambda(A) = \frac{(1 + \lambda g_\lambda)^{|A|} - 1}{\lambda}, \tag{19}$$

and $\mu_0(A) = |A|/n$. Since $1 + \lambda g_\lambda = (1 + \lambda)^{1/n}$, $\mu_\lambda(\Omega) = 1$ holds for all $\lambda > -1$.

Role of λ. With $\lambda = 0$, $R_{l,t}$ reduces to the arithmetic mean of the four patterns. $\lambda > 0$ gives a super-additive measure that emphasizes co-activation of multiple patterns, and $-1 < \lambda < 0$ gives a sub-additive measure that discounts redundant co-activation. We use $\lambda = -0.5$.

Order-statistic interpretation. With the symmetric capacity above, the integral is an ordered weighted average:

$$R_{l,t} = \sum_{i=1}^{4} w_i s_{\pi(i)}, \qquad w_i = \mu_\lambda(5 - i) - \mu_\lambda(4 - i), \tag{20}$$

where $\mu_\lambda(k)$ denotes the capacity of any set of size k. At $\lambda = -0.5$, the capacities and ascending-order weights are approximately

$$(\mu_\lambda(k))_{k=0}^{4} = (0, 0.318207, 0.585786, 0.810793, 1), \tag{21}$$

$$(w_i)_{i=1}^{4} = (0.189207, 0.225006, 0.267579, 0.318207). \tag{22}$$

Thus, larger pattern values receive more weight at the chosen setting. At $\lambda = 0$ all weights equal $1/4$; positive λ gives greater weight to smaller values and requires broader co-activation to obtain a high score.

Basic properties and numerical evaluation. The weights are nonnegative and sum to one for $\lambda > -1$, so $\min_i s_i \le R_{l,t} \le \max_i s_i$. The operator is monotone in every input and satisfies $\mathcal{A}(c, c, c, c) = c$. Equal adjacent values contribute zero differences, so tie ordering does not affect the result. The symmetric measure is permutation-invariant; it introduces no learned preference for a named pattern. Evaluation requires sorting four scalars and a weighted sum. The implementation handles the $\lambda = 0$ limit separately and clamps the final result to $[0, 1]$.

C Additional Results

C.1 Full POPE Results

Table 5 extends Table 2 with precision and recall. On LLaVA-1.5, DUST achieves the highest accuracy in all three settings. On InstructBLIP, it achieves the highest accuracy in the Popular and Adversarial settings, while OPERA is higher in the Random setting. Compared with VISTA and VTI, DUST attains higher recall at comparable precision, indicating that the accuracy gain does not come from biasing the model toward negative answers.

C.2 CHAIR Open-ended Captioning Results

Table 6 reports CHAIR results on LLaVA-1.5-7B. DUST reduces CHAIR_S from 20.00 to 18.40 while keeping CHAIR_I close to vanilla (6.23 vs. 6.18) and producing the longest captions (55.47). VISTA attains the lowest CHAIR_S (16.80), but with substantially shorter captions (43.70) and lower recall (51.00), suggesting that part of its gain comes from mentioning fewer objects. Among methods whose average caption length remains at or above 54, DUST achieves the lowest CHAIR_S. However, DUST lowers recall by 4.93 points relative to vanilla, indicating a trade-off between hallucination suppression and object coverage in open-ended generation.

C.3 Full Spatial Reasoning Results

Table 7 expands Figure 6(b) into individual metrics. DUST achieves the best scores on all Controlled-A and Controlled-B metrics, with the largest gains in set accuracy ($+26.22$ on Controlled-A and $+7.85$ on Controlled-B over vanilla), reflecting improved consistency across relation variants of the same object pair. On the natural-image subsets, DUST achieves the best COCO one-object accuracy and improves over vanilla on all four metrics, but OPERA is stronger on COCO two-object and both VG metrics.

Table 5: **Full results on POPE.** Values in parentheses denote accuracy changes relative to Vanilla. Best accuracy per setting and model is in bold.

Setting	Method	LLaVA-1.5			InstructBLIP		
		Acc.	Prec.	Rec.	Acc.	Prec.	Rec.
Random	Vanilla	82.90	92.00	72.06	80.63	81.19	79.73
	VCD	87.73 (+4.83)	91.42	72.80	80.63 (+0.00)	81.19	79.73
	ICD	83.66 (+0.76)	80.66	80.73	86.43 (+5.80)	92.01	80.73
	VISTA	83.52 (+0.62)	89.25	76.22	86.01 (+5.38)	95.20	76.73
	OPERA	85.97 (+3.07)	97.23	74.93	**89.03** (+8.40)	87.63	
	VTI	86.30 (+3.40)	97.72	74.33	86.60 (+5.97)	95.15	77.13
	DUST	**88.07** (+5.17)	96.42	78.93	86.80 (+6.17)	96.92	77.73
Popular	Vanilla	81.00	87.88	72.06	80.50	80.80	79.73
	VCD	85.38 (+4.38)	86.92	83.28	81.47 (+0.97)	82.89	79.32
	ICD	80.46 (-0.54)	76.39	88.83	82.93 (+2.43)	84.45	80.73
	VISTA	85.87 (+4.87)	94.32	76.33	84.43 (+3.93)	90.45	77.00
	OPERA	85.26 (+4.26)	94.67	74.73	80.93 (+0.43)	75.46	91.66
	VTI	85.27 (+4.27)	95.21	74.27	84.90 (+4.40)	91.32	77.13
	DUST	**86.47** (+5.47)	92.80	79.09	**85.40** (+4.90)	91.48	78.07
Adversarial	Vanilla	78.60	82.89	72.06	77.40	76.11	80.00
	VCD	80.88 (+2.28)	79.45	83.29	79.56 (+2.16)	79.67	79.37
	ICD	76.07 (-2.53)	70.77	89.47	80.87 (+3.47)	80.95	80.73
	VISTA	83.52 (+4.92)	89.25	76.22	82.20 (+4.80)	85.94	77.00
	OPERA	83.23 (+4.63)	89.84	74.93	76.33 (-1.07)	70.15	91.66
	VTI	83.50 (+4.90)	91.09	74.27	82.80 (+5.40)	86.99	77.13
	DUST	**83.70** (+5.10)	87.09	79.13	**82.92** (+5.52)	87.86	78.13

Table 6: **CHAIR results on MSCOCO captioning** (LLaVA-1.5-7B). Best in bold, second best underlined. Len. is not ranked.

Method	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	Recall↑	Len.
Vanilla	20.00	**6.18**	62.97	54.70
VCD	20.80	7.43	**63.79**	54.25
ICD	22.80	7.45	60.37	54.13
VISTA	**16.80**	7.50	51.00	43.70
OPERA	22.00	7.60	61.00	52.70
VTI	24.40	8.10	<u>63.30</u>	51.70
DUST	<u>18.40</u>	<u>6.23</u>	58.04	55.47

D LAYER-WISE PROBE AUROC

Figures 9, 10, and 11 show the full probe AUROC heatmaps summarized in Section 3.1 (Figure 2 shows a subset of metrics for LLaVA-1.5). Each row is a candidate metric computed at the last token, and each column is a decoder layer. The probing pipeline fits a separate logistic-regression classifier for each metric–layer pair and evaluates its predicted scores using validation AUROC.

The `last_token::` prefix identifies the extraction position. Multi-sample and window-based features are diagnostic probe candidates; DUST's inference gate uses only the four attention cues

Table 7: **Spatial reasoning results on What's Up** (LLaVA-1.5-7B). Best per column in bold.

Method	Controlled-A			Controlled-B			COCO		VG	
	Acc	Pair Acc	Set Acc	Acc	Pair Acc	Set Acc	One	Two	One	Two
Vanilla	80.83	63.59	38.83	82.11	64.71	36.27	63.91	61.14	40.34	38.20
VCD	61.50	39.40	0.00	73.40	22.22	0.00	51.00	56.63	41.71	37.50
ICD	43.93	14.66	3.45	36.27	0.00	0.00	43.93	40.50	33.16	29.92
VISTA	38.83	22.33	0.00	44.12	5.88	0.00	23.00	32.50	11.72	12.71
OPERA	70.39	48.28	1.72	73.04	22.22	0.00	62.44	**67.03**	**50.78**	**53.79**
VTI	62.14	29.13	3.88	62.25	42.65	23.53	63.95	51.14	37.50	24.40
DUST	**89.32**	**80.58**	**65.05**	**85.29**	**71.08**	**44.12**	**64.22**	61.36	40.78	41.58

Table 8: **Metric naming conventions in the probe registry.**

Prefix	Meaning and examples
`ll_`	Logit-lens output-distribution statistics: entropy, maximum probability, margin, perplexity.
`h_`	Hidden-state magnitude and shape: norms, variance, sparsity.
`attn_`	Attention statistics: per-head entropy, top-k mass, disagreement between heads.
`delta_`	Changes between layers in hidden states or entropy; `cum_delta_` denotes cumulative change.
`kl_`, `js_`	Divergence from the final-layer output distribution.
`mc_`	Multi-sample predictive statistics, including entropy and mutual information.
`cov_`, `eigen_`	Covariance and spectral statistics across sampled representations.
`win_`	Aggregation of statistics across a token window.

defined in Section 4.2.2. The probe classifiers and their AUROC values are not used as DUST gate weights.

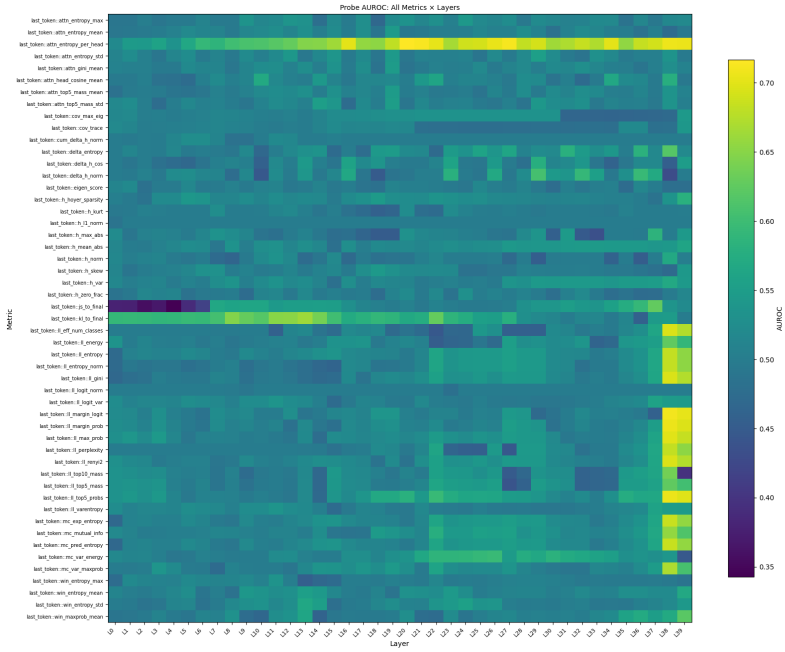


Figure 9: **Probe AUROC heatmap for LLaVA-1.5.**

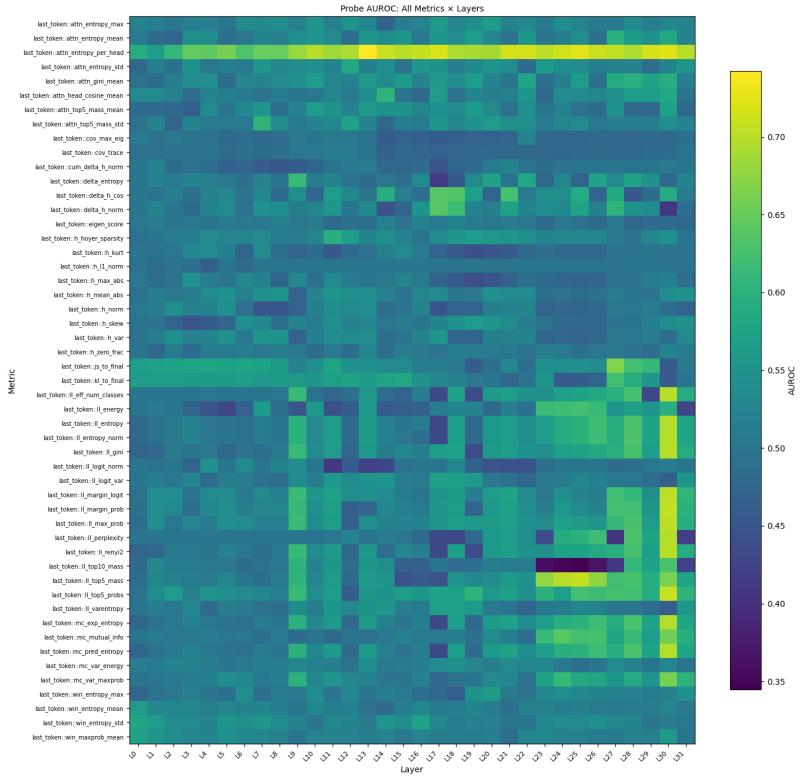


Figure 10: **Probe AUROC heatmap for InstructBLIP.**

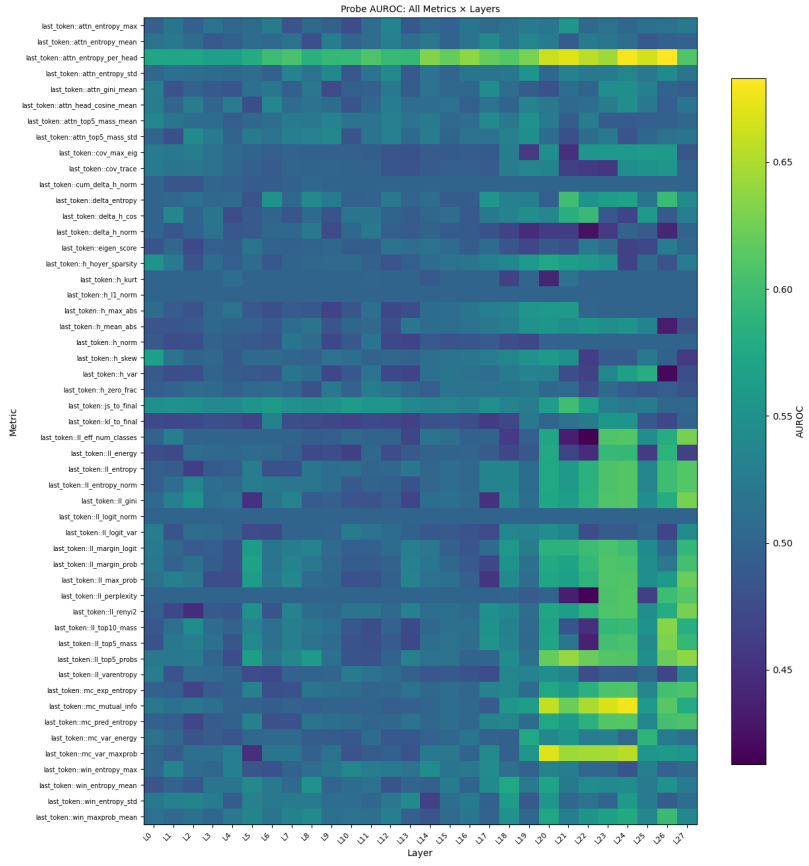


Figure 11: **Probe AUROC heatmap for Qwen2-VL-7B.**