WHEN LOOKING IS NOT ENOUGH: VISUAL ATTENTION STRUCTURE REVEALS HALLUCINATION IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have become a key interface for visual reasoning and grounded question answering, yet they remain vulnerable to visual hallucinations, where generated responses contradict image content or mention nonexistent objects. A central challenge is that hallucination can coexist with substantial visual attention: the model may assign considerable attention mass to image tokens while internally drifting toward an incorrect answer. In this paper, we find that the high-frequency structure of visual attention, measured by layer-wise Laplacian energy, helps select intermediate layers with contrasting preferences for hallucinated and correct answers. Building on this finding, we propose **LaSCD** (**La**placian-**S**pectral **C**ontrastive **D**ecoding), a training-free decoding strategy that selects informative layers via Laplacian energy and remaps next-token logits in closed form. Experiments on hallucination and general multimodal benchmarks show that LaSCD consistently reduces hallucination while preserving general capabilities, highlighting its potential as a faithful decoding paradigm.

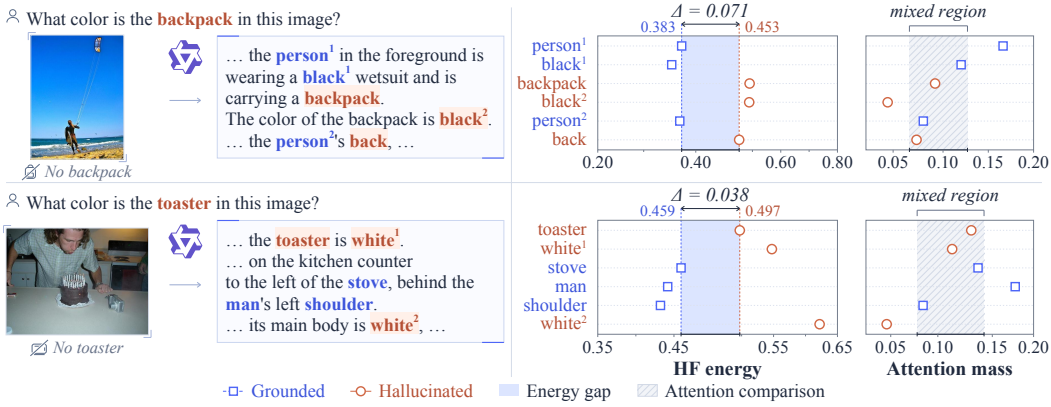

Figure 1: Qualitative comparison of high-frequency (HF) energy and attention mass in responses generated by Qwen3-VL-8B-Instruct. The model is asked to describe the color of an object that is entirely absent from the image. Grounded and hallucinated word occurrences are marked in blue (□) and orange (○), respectively. Hallucinated words exhibit higher HF energy in these examples; for instance, the hallucinated `backpack` has an energy of 0.490, compared with 0.383 for the grounded `person`. HF energy clearly separates hallucinated from grounded words.

1 INTRODUCTION

Multimodal large language models (MLLMs) have become a central interface for visual reasoning, grounded dialogue, and open-ended visual question answering (Alayrac et al., 2022; Liu et al., 2024c; Bai et al., 2023; Liu et al., 2024b). Despite this progress, they still suffer from visual hallucinations: generated descriptions or answers can contradict the image, mention nonexistent objects, or make unsupported visual judgments (Huang et al., 2024b; Zheng et al., 2024; Lyu et al., 2025). This failure is especially concerning in safety sensitive applications such as healthcare and autonomous driving (Lin et al., 2024b; Ding et al., 2024). Existing mitigation methods address hallucination from different angles, including retrieval augmentation (Qu et al., 2024; Yang et al., 2024), fine-tuning or preference optimization (Yu et al., 2024a; Liu et al., 2024a), attention intervention (Huang et al., 2024a), and contrastive decoding (CD) (Leng et al., 2024; Li et al., 2023b; Shi et al., 2024). These methods improve faithfulness in many cases and motivate a complementary question: *what internal visual signal can help identify layers whose predictions are useful for mitigating hallucination?*

A common intuition links hallucination to insufficient use of visual evidence. This intuition has motivated methods that increase, reallocate, or intervene on visual attention during decoding (Huang

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.

et al., 2024a; Liu et al., 2024d; Zou et al., 2025). However, total visual-attention mass provides a limited summary of how that evidence is used (Sec. 3.2). The model can assign substantial attention to visual tokens while favoring an incorrect answer, motivating an examination of the spatial structure of that attention. This distinction matters for both analysis and mitigation: magnitude-based signals can activate at the wrong layer where the model is already visually engaged but internally misaligned.

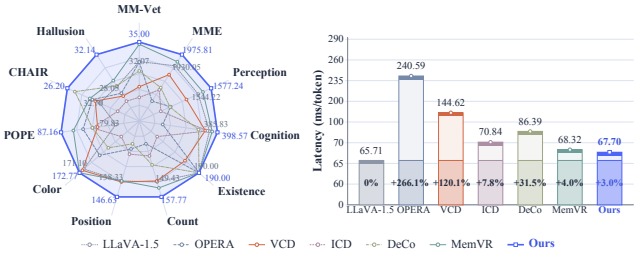

Figure 2: Performance and latency comparison of SOTA.

We present an empirical analysis of visual-attention *magnitude* and *high-frequency structure* as hallucination detecting signals. Specifically, we compute the norm of a discrete Laplacian response over local neighborhoods of visual tokens in the last-query attention slice. Our analysis reveals two complementary phenomena (Fig. 3, 4). First, hallucinated outputs can occur even with high visual-attention mass, indicating that strong visual attention does not necessarily imply visual grounding. Second, hallucinated tokens are typically associated with high Laplacian energy, whereas grounded tokens are associated with low energy. These findings motivate using the *spatial variation* of visual attention to guide hallucination detection and mitigation.

Motivated by this insight, we propose **LaSCD** (**La**placian-**S**pectral **C**ontrastive **D**ecoding), a training-free decoding strategy for MLLMs. At each decoding step, LaSCD scores candidate layers using high-frequency energy, selects a hallucination-associated peak layer and a ground-truth-associated low-energy layer, and then remaps next-token logits in closed form. The core operation contrasts the final logits against the peak-layer logits, suppressing tokens that are already over-supported at the layer where hallucination starts to form. An optional correction can further use the low-energy layer when validation suggests a benefit. Importantly, LaSCD uses one regular autoregressive forward and only adds lightweight LM-head evaluations and attention-side energy computation, avoiding the second full forward typical of two-view CD pipelines.

We evaluate LaSCD on standard hallucination and general multimodal benchmarks, where LaSCD achieves state-of-the-art performance: reducing hallucination while preserving general capabilities where most existing methods fail. Results support effectiveness of the proposed structural signal.

Our contributions are summarized as follows:

- We identify a layer-wise empirical phenomenon in MLLMs: visual-attention mass can remain high during hallucination while Laplacian energy over the visual-attention structure reveals both hallucination onset and transient ground-truth recovery.
- We introduce LaSCD, a training-free decoding method that converts this finding into a simple closed-form logit remapping, contrasting the final prediction against the energy-peak layer and optionally correcting with the low-energy layer.
- We validate LaSCD across hallucination and general multimodal benchmarks, showing improved faithfulness with modest inference overhead and ablation evidence that supports the proposed visual-attention structure signal.

2 Related Work

Multimodal LMs and hallucination. Multimodal large language models (MLLMs) have advanced rapidly, building on vision–language pretraining and instruction tuning (Alayrac et al., 2022; Liu et al., 2024c; Bai et al., 2023; Liu et al., 2024b). Despite strong task performance, they remain prone to *visual hallucination*, producing text that is inconsistent with the image, which has motivated a growing line of analysis and mitigation (Yin et al., 2023; Bai et al., 2024).

Mitigation strategies. A first family of methods improves alignment or rewrites model outputs with extra supervision: hallucination-oriented fine-tuning or detection datasets, learned revisors, or preference optimization (Gunjal et al., 2024; Zhou et al., 2024; Yu et al., 2024a), requiring additional data or optimization. A complementary line stays *training-free* at test time. **Attention-intervention** methods directly manipulate attention maps or apply decoding-time penalties (Huang et al., 2024a; Zhang et al., 2024; Xing et al., 2024; Liu et al., 2024d; Xu et al., 2025; Yin et al.,

Table 1: Comparison of decoding-time hallucination mitigation methods.

Methods	No Second Input View	Hallucination Signal	Visual-Structure Aware	Modified Component(s)	Decoding Paradigm
DoLa Chuang et al. (2023)	✓	✓	✗	Logits	Contrast
OPERA Huang et al. (2024a)	✓	✗	✗	Attention matrix	Att-intervention
PAI Liu et al. (2024d)	✗	✗	✗	Attention matrix	Att-intervention
VCD Leng et al. (2024)	✗	✗	✗	Visual input, logits	Contrast
ICD Wang et al. (2024a)	✗	✗	✗	Textual input, logits	Contrast
SID Huo et al. (2024)	✗	✗	✗	Textual input, logits	Contrast
DeCo Wang et al. (2025)	✓	✗	✗	Logits	Contrast
MemVR Zou et al. (2025)	✓	✓	✗	Hidden states	Visual-retracing
AIR Zhu et al. (2026)	✓	✓	✗	Hidden states	Visual-retracing
DMAS Yin et al. (2026)	✓	✗	✗	Attention matrix	Att-intervention
LaSCD (ours)	✓	✓	✓	Logits	**Spectral-contrast**

2026). Their runtime cost depends on attention access and any additional decoding steps. A second **contrastive-decoding (CD)** family compares logits across layers or input conditions to down-weight spurious continuations, including layer-contrast methods in language models (Chuang et al., 2023) and visual- or instruction-contrast counterparts in MLLMs such as VCD/ICD (Leng et al., 2024; Wang et al., 2024a; Chen et al., 2024; Neo & Chen, 2024; Wang et al., 2026; Song et al., 2026; Ren et al., 2026). These approaches retain the pretrained backbone and differ in how they construct the contrastive distribution; *two-view* realizations incur an additional end-to-end forward pass. Recently MemVR (Zou et al., 2025) and AIR (Zhu et al., 2026) have explored **inference-time visual reinforcement**, explicitly re-integrating visual evidence into the language decoder to address visual amnesia or shifts toward textual information. Specifically, MemVR selects high-uncertainty layers using entropy while AIR employs *Optimal Transport* to measure alignment between hidden states and patch embeddings. However, these methods typically rely on coarse-grained statistical metrics to approximate the model's hallucination state. Such aggregate statistics discard the structural dynamics, which are critical for localizing hallucination onset.

Positioning of LaSCD. LaSCD is training-free and operates entirely at decoding time. It scores *layer-wise* visual-attention variation using Laplacian energy and then applies a *closed-form* logit remapping. Following the CD *principle*, it contrasts the final head output against an intermediate one selected within a *single* autoregressive forward. Relative to layer-wise decoders such as DoLa (Chuang et al., 2023) and DeCo (Wang et al., 2025), the distinguishing choice is to select layers using a visual-attention score and combine their logits through energy-peak contrast and a minimum-energy correction. Table 1 compares representative decoding-time approaches.

3 Empirical Analysis

3.1 Preliminaries

Multimodal large language models. MLLMs pair a pretrained vision encoder with a decoder-only language model through a lightweight projector. Given an image or video $\mathcal{I}$ and a text prompt $\mathcal{P}$, the vision encoder maps $\mathcal{I}$ to visual tokens $\mathbf{V} \in \mathbb{R}^{N_v \times d}$, which are projected into the language embedding space and concatenated with text tokens $\mathbf{T} \in \mathbb{R}^{N_t \times d}$ to form a single input sequence $\mathbf{X} = [\mathbf{T}_{\text{pre}}, \mathbf{V}, \mathbf{T}_{\text{post}}]$ of length $N = N_v + N_t$. We denote by $\mathcal{I}_v \subset \{1, \dots, N_v\}$ the index set of visual tokens in $\mathbf{X}$. At each decoding step, an L-layer Transformer decoder consumes $\mathbf{X}$ together with the previously generated tokens. At layer l, the last-position (i.e. currently generated) query attends to all N keys; we write $\mathbf{A}^{(l,h)} \in \mathbb{R}^N$ for the attention distribution of head $h \in \{1, \dots, H\}$ over key positions, and obtain the next-token logits from the final-layer hidden state via an unembedding head, $\mathbf{z}^{(L)} = \text{LMHead}(\mathbf{h}^{(L)})$, from which the next token is sampled.

Multi-modal Hallucination. Prior work mitigates MLLM hallucination by modifying attention or reinforcing visual evidence during decoding (Huang et al., 2024a; Liu et al., 2024d; Zou et al., 2025). To study signals available for layer selection, we examine the visual slice of the last-query attention. Concretely, we denote the head-averaged attention from the current token to all visual keys $\boldsymbol{\alpha}_{\text{vis}}^{(l)}$, together with its scalar ℓ_1 mass, by:

$$\boldsymbol{\alpha}_{\text{vis}}^{(l)} \triangleq \frac{1}{H} \sum_{h=1}^{H} \mathbf{A}^{(l,h)}[\mathcal{I}_v] \in \mathbb{R}^{N_v}, \qquad m^{(l)} \triangleq \left\| \boldsymbol{\alpha}_{\text{vis}}^{(l)} \right\|_1 = \sum_{i \in \mathcal{I}_v} \alpha_{\text{vis},i}^{(l)} \in \mathbb{R}. \tag{1}$$

Many prior methods operationalize $\boldsymbol{\alpha}_{\text{vis}}^{(l)}$ in two ways: (i) as a *trigger* that decides *when* to activate a mitigation step (e.g., when $m^{(l)}$ appears small), or (ii) as an *edit target*—reallocating, suppressing, or

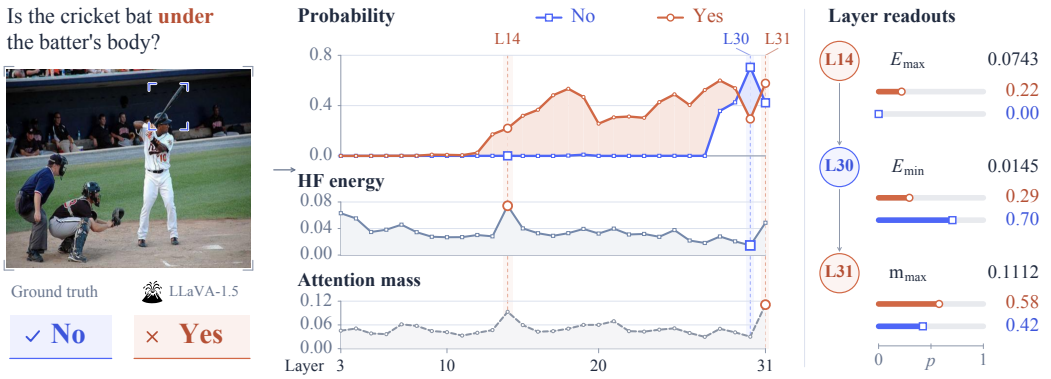

Figure 3: **Three views of the same decoding step with an incorrect final answer.** LLaVA-1.5 is asked *"Is the cricket bat under the batter's body?"*; ground truth is `No`, but the model answers `Yes` with $p(\texttt{Yes}) = 0.58$. Using the *same* attention patterns, for each layer we plot: **(a)** probabilities of `Yes` and `No` from the hidden state at that layer; **(b)** high-frequency energy $E^{(l)}$ of $\boldsymbol{\alpha}_{\text{vis}}^{(l)}$ from Eq. 3 and **(c)** visual attention mass $m^{(l)}$ from Eq. 1. During the decoding stage, the layer that maximizes $E^{(l)}$ coincides with a rise in the hallucinated token's probability, while the layer that minimizes $E^{(l)}$ is associated with a higher probability for the ground-truth token. More cases is available at Appendix D.

rewiring attention over keys during decoding so that subsequent hidden states and logits are steered toward visually grounded predictions. The empirical analysis below highlights scenarios where visual-attention magnitude offers limited signal for identifying hallucination.

3.2 Is Visual-Attention Magnitude a Solid Hallucination Signal?

A common assumption implicit in Sec. 3.1 is that, at the layer that commits to the next token, hallucinated outputs are associated with *lower* visual-attention mass $m^{(l)}$ and confidently correct outputs with *higher* $m^{(l)}$. We revisit this assumption through a representative case (Figure 3) and a quantitative analysis (Figure 4). The two observations below suggest that visual-attention magnitude alone may not be sufficient as a hallucination indicator in some scenarios.

Confident hallucination coincides with high visual attention magnitude. Two observations in Fig. 3 illustrate the limitations of total mass as a hallucination signal. First, the hallucination does not arise as a last-layer artifact: the logit of *Yes* surpasses that of *No* in the intermediate layers and reaches a large margin well prior to the final output, indicating that the model's commitment to the erroneous response is established and sustained across a substantial portion of the forward pass. Second, $m^{(l)}$ remains substantially high across the shallow and intermediate layers throughout the onset of hallucination and ultimately attains its global maximum precisely at the layer responsible for emitting the hallucinated token.

In this example, high visual-attention magnitude coexists with an incorrect answer preference, and $m^{(l)}$ reaches its maximum at the layer that emits the hallucinated token. Since $m^{(l)}$ aggregates over visual tokens without regard to their spatial arrangement, this observation motivates examining the structure of the attention distribution alongside its total mass.

Observation 1

High visual-attention mass can coexist with an hallucinated answer preference.

3.3 Laplacian Energy and Layer-wise Answer Preferences

While attention magnitude fails to isolate these internal dynamics, we further investigate the distribution of attention over visual tokens. Inspired by attention fluctuations observed in unimodal generation (Qi et al., 2026), we summarize spatial structure using *Laplacian energy*. Let $\boldsymbol{\alpha}_{\text{vis}}^{(l,h)} \triangleq \mathbf{A}^{(l,h)}[\mathcal{I}_v] \in \mathbb{R}^{N_v}$ denote the per-head visual slice. Our default operator is the discrete Laplacian, $\mathcal{K} = \Delta_{\mathcal{N}}$, defined on an undirected token neighborhood $\mathcal{N}(i)$ by

$$[\Delta_{\mathcal{N}}\mathbf{x}]_i = \sum_{j\in\mathcal{N}(i)} (x_i - x_j) + b_i x_i, \tag{2}$$

where b_i counts fixed zero-valued exterior neighbors. Sequence and grid layouts are represented by their respective neighborhoods $\mathcal{N}$ under this common definition (Appendix B.4). For a chosen operator $\mathcal{K}$, fixed across heads and candidate layers for each input, we compute

$$E^{(l)}(\mathcal{K}) = \frac{1}{H}\sum_{h=1}^{H} \left\| \mathcal{K}\boldsymbol{\alpha}_{\text{vis}}^{(l,h)} \right\|_2, \qquad E^{(l)} \equiv E^{(l)}(\Delta_{\mathcal{N}}). \tag{3}$$

The default Laplacian response emphasizes higher-frequency components relative to the neighborhood. We also evaluated Sobel-style first differences ($E^{(l)}(\mathcal{K}_{\nabla})$) and Laplacian-of-Gaussian filters ($E^{(l)}(\mathcal{K}_{\text{LoG}})$); all choices reproduced the qualitative behavior in Fig. 3.

Laplacian energy better identifies grounded and hallucinated preferences. In Figure 3, the energy peak precedes a rise in the hallucinated token's probability while the minimum-energy layer favors the grounded answer. Figure 4(a) extends the analysis across hallucination cases: maximum-energy layers yield higher hallucination rates than minimum-mass layers for both Qwen and LLaVA. These increases show that energy-based detection more consistently captures hallucinated token preferences in the analyzed subsets. Two examples in Figure 1 also show higher energy for hallucinated object and attribute mentions than for grounded mentions, whereas their attention masses overlap. These complementary observations support Laplacian energy as a more informative indicator of token grounding than attention mass and motivate its use for layer selection.

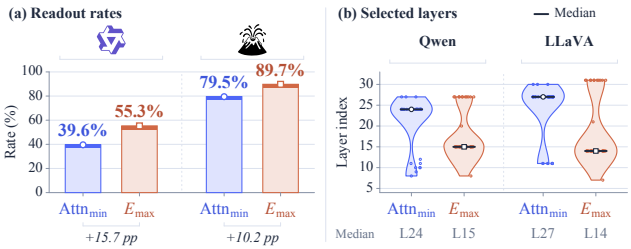

Figure 4: Hallucinated-token readouts on MME-Hall and CHAIR subsets with hallucination final answers. Panel (a) compares hallucination rates at layers selected by minimum visual-attention mass (Attn_{min}) and maximum Laplacian energy (E_{max}) for Qwen and LLaVA. Maximum-energy selection yields higher hallucination rates for both models.

Observation 2

Energy signals separate hallucinated from grounded token preferences more sharply.

4 Methodology

4.1 Overview

Sec. 3.2–3.3 show that Laplacian energy $E^{(\ell)}$ complements visual-attention *magnitude* as a signal for selecting layers with contrasting answer preferences.

Motivated by these observations, we propose **LaSCD** (**La**placian-**S**pectral **C**ontrastive **D**ecoding), a training-free decoding strategy. LaSCD scores layers with $E^{(\ell)}$ (Sec. 4.2) then remaps next-token logits: (i) contrast against the peak-layer preference relative to $\mathbf{z}^{(L)}$ (Sec. 4.3); (ii) a confidence-weighted correction from ℓ_{min} (Sec. 4.4). Section 4.5 gives the closed-form logits.

4.2 Layer Selection

We first evaluated $E^{(\ell)}$ on a chosen set of decoder layers $\mathcal{L}$. At each step we take the last-query row of $\mathbf{A}^{(\ell,h)}$, restrict to visual keys when analyzing vision tokens, map the slice to a grid when applicable, and compute $E^{(\ell)}$ via the Laplacian response (Eq. 3); alternative spatial layouts follow the same principle as in the empirical section. We then select

$$\ell_{\text{peak}} \in \arg\max_{\ell\in\mathcal{L}} E^{(\ell)}, \qquad \ell_{\text{gt}} \in \arg\min_{\ell\in\mathcal{L}} E^{(\ell)}, \tag{4}$$

during the forward process. ℓ_{peak} targets the onset of high-curvature, hallucination-associated attention structure while ℓ_{gt} targets a depth that often briefly aligns with the correct token.

4.3 Contrastive Logits (LaSCD-α)

Intermediate logits at ℓ_{peak} tend to lock in the eventual error with a large margin, which the final layer inherits; we therefore contrast the final logits against that depth while keeping $\mathbf{z}^{(L)}$ as the anchor:

$$\tilde{\mathbf{z}} \;=\; (1+\alpha)\,\mathbf{z}^{(L)} \;-\; \alpha\,\mathbf{z}^{(\ell_{\text{peak}})}, \qquad \alpha \geq 0, \tag{5}$$

where $\mathbf{z}^{(\ell_{\text{peak}})}$ is the LM-head output at ℓ_{peak}. When $\mathbf{z}^{(\ell_{\text{peak}})} \approx \mathbf{z}^{(L)}$, Eq. 5 leaves benign trajectories largely unchanged. We apply top-K masking on the *composed* logits, following standard contrastive-decoding practice.

4.4 Decoding Correction (LaSCD-β)

The minimum-energy depth ℓ_{gt} can carry a complementary preference aligned with the ground-truth token (Sec. 3.3). To add a small correction $\propto \mathbf{z}^{(\ell_{\text{gt}})}$ without letting fixed scaling amplify mismatched logit scales across depth, we gate the strength by the peak softmax mass of that layer's distribution:

$$\pi_{\max}^{(\ell)} \;\triangleq\; \max_{v\in\mathcal{V}} \left[\text{softmax}(\mathbf{z}^{(\ell)})\right]_v, \qquad \beta_{\text{eff}} \;=\; \beta \cdot \pi_{\max}^{(\ell_{\text{gt}})}, \quad \beta \geq 0. \tag{6}$$

4.5 Closed-Form of LaSCD

LaSCD-α and LaSCD-β combine additively:

$$\tilde{\mathbf{z}} \;=\; (1+\alpha)\,\mathbf{z}^{(L)} \;-\; \alpha\,\mathbf{z}^{(\ell_{\text{peak}})} \;+\; \beta_{\text{eff}}\,\mathbf{z}^{(\ell_{\text{gt}})}. \tag{7}$$

Complexity Analysis. At each autoregressive step we use *one* forward F to compute $E^{(\ell)}$ on $\mathcal{L}$, pick ℓ_{peak} and ℓ_{gt}, and form $\tilde{\mathbf{z}}$ in Eq. 7; the overhead beyond F is $\mathcal{O}(|\mathcal{L}|\,H\,N_v)$ for the Laplacian energy and $\mathcal{O}(|\mathcal{V}|)$ for the logit mix. By contrast, *existing* methods in Table 2 can incur (i) a *second* full forward ($2F$) when the scheme requires a second pass through the stack, (ii) a *per-probed-layer* cost that multiplies $|\mathcal{L}|$ with a full $|\mathcal{V}|$-dimensional logit/entropy readout, or (iii) a post-F re-weighting that scales with model hidden dimension as $\mathcal{O}(N_v d)$. LaSCD avoids the duplicated backbone pass, keeps a *single* F, and confines the nontrivial work to the attention–vision slice in $\mathcal{O}(|\mathcal{L}|\,H\,N_v)$ with only a *bounded* $\mathcal{O}(d|\mathcal{V}|)$ logit merge in Eq. 7, which sidesteps the $d|\mathcal{L}||\mathcal{V}|$ and $N_v d$ patterns above and is typically more favorable in large-vocab settings.

Table 2: Complexity comparison of different methods; F: one autoregressive forward, H: number of attention heads, N_v: visual positions, d: hidden dim, $\mathcal{L}$: layer subsets, and $|\mathcal{V}|$: the vocabulary size.

	VCD	DeCo	MemVR	LaSCD (ours)
Complexity	$\mathcal{O}(2F + d\|\mathcal{V}\|)$	$\mathcal{O}(F + d\|\mathcal{L}\|\,\|\mathcal{V}\|)$	$\mathcal{O}(F + d\|\mathcal{L}\|\,\|\mathcal{V}\| + N_v d)$	$\mathcal{O}(F + \|\mathcal{L}\|\,H\,N_v + d\|\mathcal{V}\|)$

Theoretical Analysis. For normalized attention $\mathbf{p}$, define $Q_{\mathcal{N}}(\mathbf{p}) = \|\Delta_{\mathcal{N}}\mathbf{p}\|_2^2$. Appendix A derives its spectral weighting and boundary response under a prescribed neighborhood and proves the following identity.

Theorem 1. *Let $\boldsymbol{\pi}$ be a normalized attention vector and $\boldsymbol{\eta}$ a zero-mean perturbation with finite covariance Σ_η, preserving nonnegativity and unit mass almost surely. For fixed $\Delta_{\mathcal{N}}$,*

$$\mathbb{E}[Q_{\mathcal{N}}(\boldsymbol{\pi}+\boldsymbol{\eta})] = Q_{\mathcal{N}}(\boldsymbol{\pi}) + \text{Tr}\big(\Delta_{\mathcal{N}}^{\top}\Delta_{\mathcal{N}}\Sigma_\eta\big)\,. \tag{8}$$

The expected squared response increases strictly whenever the trace term is positive.

Remark. The identity characterizes sensitivity to attention perturbations for a fixed neighborhood. Appendix A.4 relates this normalized squared response to the raw, unsquared decoding score, whose association with answer preferences is examined in Sec. 3.3.

5 Experiment

5.1 Experiment Setup

Baselines and models. We evaluate five backbones: InstructBLIP (Li et al., 2023a), LLaVA-1.5-7B (Liu et al., 2023), Qwen-VL-Chat (Bai et al., 2023), GLM-4V-9B (Glm et al., 2024), and Qwen3-VL-8B-Instruct (Bai et al., 2025). Video backbones include Video-LLaVA-7B (Lin et al., 2024a)

Table 3: Performance evaluation on POPE benchmark. The best results are in green.

Evaluation	Methods	Random		Popular		Adversarial		Average	
		Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑
LLaVA-1.5	Regular	83.49 ↑0.0	82.28 ↑0.0	79.98 ↑0.0	79.34 ↑0.0	76.03 ↑0.0	76.26 ↑0.0	79.83 ↑0.0	79.29 ↑0.0
	OPERA (Huang et al., 2024a)	87.53 ↑4.0	86.45 ↑4.2	84.21 ↑4.2	83.50 ↑4.2	80.88 ↑4.9	80.69 ↑4.4	84.21 ↑4.4	83.55 ↑4.3
	DoLa (Chuang et al., 2023)	85.78 ↑2.2	84.69 ↑2.4	80.75 ↑0.7	81.11 ↑1.7	77.32 ↑1.3	76.66 ↑0.4	81.28 ↑1.4	80.82 ↑1.5
	VCD (Leng et al., 2024)	86.84 ↑3.4	86.83 ↑4.6	82.65 ↑2.7	83.37 ↑4.0	77.31 ↑1.3	79.28 ↑3.0	82.27 ↑2.4	83.16 ↑3.9
	ICD (Wang et al., 2024a)	84.87 ↑1.4	83.27 ↑1.0	82.93 ↑3.0	81.45 ↑2.1	81.07 ↑5.0	79.96 ↑3.7	82.96 ↑3.1	81.56 ↑2.3
	SID (Huo et al., 2024)	87.86 ↑4.4	85.73 ↑3.4	83.79 ↑3.8	85.42 ↑6.1	82.54 ↑6.5	81.98 ↑5.7	84.73 ↑4.9	84.37 ↑5.0
	DeCo (Wang et al., 2025)	86.66 ↑3.2	89.15 ↑6.9	84.63 ↑4.7	85.84 ↑6.5	77.63 ↑1.6	80.64 ↑4.4	82.97 ↑3.1	85.21 ↑5.9
	MemVR (Zou et al., 2025)	88.50 ↑5.0	87.34 ↑5.0	86.07 ↑6.1	85.60 ↑6.2	81.97 ↑5.9	81.33 ↑5.0	85.51 ↑5.6	84.75 ↑5.4
	ASCD (Wang et al., 2026)	89.48 ↑6.0	89.09 ↑6.8	87.20 ↑7.2	86.69 ↑7.4	82.90 ↑6.9	82.97 ↑6.7	86.53 ↑6.7	86.25 ↑7.0
	NoLan (Ren et al., 2026)	87.11 ↑3.6	86.60 ↑4.3	85.81 ↑5.8	85.17 ↑5.8	83.83 ↑7.8	82.58 ↑6.3	85.58 ↑5.8	84.78 ↑5.5
	AIR (Zhu et al., 2026)	89.00 ↑5.5	88.20 ↑5.9	87.10 ↑7.1	86.40 ↑7.1	83.90 ↑7.9	83.60 ↑7.3	86.67 ↑6.8	86.07 ↑6.8
	DMAS (Yin et al., 2026)	90.03 ↑6.5	90.02 ↑7.7	87.33 ↑7.4	87.03 ↑7.7	83.07 ↑7.0	83.33 ↑7.1	86.81 ↑7.0	86.79 ↑7.5
	Ours	90.57 ↑7.0	90.86 ↑8.0	87.45 ↑7.5	87.40 ↑8.0	83.46 ↑7.4	82.60 ↑6.3	87.16 ↑7.3	86.95 ↑7.6
InstructBLIP	Regular	80.42 ↑0.0	80.94 ↑0.0	76.09 ↑0.0	77.65 ↑0.0	72.37 ↑0.0	75.42 ↑0.0	76.29 ↑0.0	78.00 ↑0.0
	OPERA (Huang et al., 2024a)	84.57 ↑4.2	84.69 ↑3.8	78.52 ↑2.4	81.02 ↑3.4	75.60 ↑3.2	78.37 ↑3.0	79.56 ↑3.3	81.36 ↑3.4
	DoLa (Chuang et al., 2023)	83.01 ↑2.6	83.04 ↑2.1	78.97 ↑2.9	79.83 ↑2.2	74.76 ↑2.4	76.58 ↑1.2	78.91 ↑2.6	79.82 ↑1.8
	VCD (Leng et al., 2024)	84.11 ↑3.7	84.13 ↑3.2	79.64 ↑3.6	80.80 ↑3.2	76.42 ↑4.1	78.03 ↑2.6	80.06 ↑3.8	80.99 ↑3.0
	ICD (Wang et al., 2024a)	82.57 ↑2.2	83.32 ↑2.4	78.23 ↑2.1	79.12 ↑1.5	75.28 ↑2.9	76.24 ↑0.8	78.69 ↑2.4	79.56 ↑1.6
	SID (Huo et al., 2024)	86.56 ↑6.1	85.94 ↑5.0	80.26 ↑4.2	81.75 ↑4.1	77.64 ↑5.3	80.41 ↑5.0	81.49 ↑5.2	82.70 ↑4.7
	Ours	90.02 ↑9.6	89.75 ↑8.8	83.56 ↑7.4	83.87 ↑6.2	81.03 ↑8.6	81.83 ↑6.4	84.87 ↑8.5	85.15 ↑7.1

Table 4: HallusionBench evaluation results.

Methods	fACC ↑	qACC ↑	*easy*A ↑	*hard*A ↑	aACC ↑
LLaVA-1.5	17.9 ↑0.0	8.13 ↑0.0	36.0 ↑0.0	36.7 ↑0.0	41.5 ↑0.0
+ OPERA	16.2 ↓1.7	5.49 ↓2.6	37.6 ↑1.6	35.4 ↓1.3	41.2 ↓0.3
+ ICD	13.9 ↓4.0	8.35 ↑0.2	36.9 ↑0.9	33.5 ↓3.2	38.2 ↓3.3
+ VCD	13.9 ↓4.0	11.4 ↑3.3	33.0 ↓3.0	34.7 ↓2.0	41.1 ↓0.4
+ DeCo	15.3 ↓2.6	8.13 ↑0.0	42.1 ↑6.1	33.8 ↓2.9	41.3 ↓0.2
+ MemVR	17.9 ↑0.0	9.01 ↑0.9	36.9 ↑0.9	37.7 ↑1.0	42.5 ↑1.0
+ Ours	15.8 ↓2.1	13.8 ↑5.7	41.5 ↑5.5	43.4 ↑6.7	46.2 ↑4.7
Qwen-VL	19.9 ↑0.0	12.3 ↑0.0	39.3 ↑0.0	32.0 ↑0.0	41.8 ↑0.0
+ DeCo	20.2 ↑0.3	14.0 ↑1.7	42.8 ↑3.5	34.6 ↑2.6	43.6 ↑1.8
+ Ours	22.2 ↑2.3	13.8 ↑1.5	43.0 ↑3.7	35.5 ↑3.5	44.9 ↑3.1

Table 5: LLaVA-Bench evaluation results.

Method	LLaVA-Bench (in-the-wild)			
	Average↑	All_1↑	All_2↑	All_3↑
LLaVA-1.5	45.1 ↑0.0	25.4 ↑0.0	87.8 ↑0.0	22.3 ↑0.0
+ DeCo	45.6 ↑0.5	25.9 ↑0.5	87.9 ↑0.1	23.0 ↑0.7
+ Ours	45.9 ↑0.8	26.5 ↑1.1	88.0 ↑0.2	23.3 ↑1.0
Qwen-VL	44.5 ↑0.0	25.0 ↑0.0	86.8 ↑0.0	21.7 ↑0.0
+ DeCo	44.7 ↑0.2	25.6 ↑0.6	85.8 ↓1.0	22.8 ↑1.1
+ Ours	44.9 ↑0.4	26.1 ↑1.1	86.3 ↓0.5	22.5 ↑0.8
GLM-4V	49.1 ↑0.0	33.2 ↑0.0	85.8 ↑0.0	28.5 ↑0.0
+ DeCo	48.3 ↓0.8	31.8 ↓1.4	85.8 ↑0.0	27.3 ↓1.2
+ Ours	49.3 ↑0.2	32.9 ↓0.3	86.5 ↑0.7	28.5 ↑0.0

and LLaVA-Next-Video-7B-DPO (Liu et al., 2024b). We compare with training-free hallucination mitigation methods: OPERA (Huang et al., 2024a), DoLa (Chuang et al., 2023), VCD (Leng et al., 2024), ICD (Wang et al., 2024a), SID (Huo et al., 2024), DeCo (Wang et al., 2025), MemVR (Zou et al., 2025), ASCD (Wang et al., 2026), NoLan (Ren et al., 2026), AIR (Zhu et al., 2026), DMAS (Yin et al., 2026) and Self-PEP (Wang et al., 2024b).

Datasets and Evaluation. To rigorously assess the effectiveness of our proposed method, we conduct a comprehensive set of experiments across POPE benchmark (Li et al., 2023c), CHAIR (Rohrbach et al., 2018), MME (Fu et al., 2023), MM-Vet (Yu et al., 2024b), LLaVA-Bench (in-the-wild) (Liu et al., 2024c), HallusionBench (Guan et al., 2024) and VideoHallucer benchmark (Wang et al., 2024b).

Implementation Details. Usually, we set $\alpha = 0.1$, $\beta = 0.2$ and $\mathcal{L} = \{8, 9, \ldots, 29\}$. Baseline configurations follow the original papers. More details are in Appendix B.

5.2 Evaluations on Hallucination Benchmarks

We evaluate hallucination on CHAIR, POPE, and HallusionBench, and further test generalization on video understanding; results are in Table 6, Table 3, Table 4 and Table 9. In the POPE evaluation, LaSCD is effective across two backbones, clearly surpassing prior SOTAs, with an average accuracy increase of up to +7.9% and an F1-score increase of up to +7.3% under the Random, Popular, and Adversarial settings. On CHAIR dataset, LaSCD improves performance by 19.8% and 34.7% over vanilla LLaVA-1.5 and Qwen-VL-Chat. Table 4 shows the HallusionBench evaluation results, where LaSCD consistently yields the best results among all comparators, including hardaACC and aACC. On video hallucination, LaSCD generalizes well and outperforms baselines by large margins, demonstrating strong robustness beyond image understanding benchmarks.

5.3 Evaluations on General Capability Benchmarks

We evaluate LaSCD on general-capability benchmarks, including MME, MM-Vet, and LLaVA-Bench. We tested three different MLLM backbones. On LLaVA-Bench (Table 5), LaSCD consistently surpasses the compared baselines. On MME subsets that target object- and attribute-level hallucination together with perception- and cognition-oriented items (Table 10), LaSCD delivers uniform improvements at both granularities: the aggregate MME-Hall score increases by 26 and 21 point for the

Table 6: CHAIR results with up to 1024 output tokens. The best results are in green.

Methods	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	Average ↓	Len	Recall ↑
LLaVA-1.5	50.0 ↓0.0	15.4 ↓0.0	32.7 ↓0.0	100.6	77.1 ↑0.0
+ OPERA	47.8 ↓2.2	14.6 ↓0.8	31.2 ↓0.5	98.6	76.8 ↓0.3
+ VCD	48.6 ↓1.4	14.9 ↓0.5	31.8 ↓0.5	100.4	77.3 ↑0.2
+ ICD	56.2 ↑6.2	16.3 ↑0.9	36.3 ↑3.6	103.4	16.3 ↓60.
+ DeCo	42.8 ↓7.2	12.6 ↓2.8	27.7 ↓5.0	103.4	76.7 ↓0.4
+ MemVR	47.6 ↓2.4	13.8 ↓1.6	30.7 ↓2.0	101.5	79.6 ↑2.5
+ Ours	40.6 ↓9.4	11.8 ↓2.4	26.2 ↓6.5	100.0	77.6 ↑0.5
Qwen-VL	6.80 ↑0.0	5.30 ↑0.0	6.05 ↑0.0	17.6	53.4 ↑0.0
+ OPERA	-	-	-	-	-
+ VCD	13.0 ↑6.2	12.3 ↑7.0	12.6 ↑6.5	115.	47.9 ↓5.5
+ ICD	18.4 ↑11.	14.3 ↑9.0	16.3 ↑10.	48.1	37.6 ↓15.
+ DeCo	12.2 ↑5.4	6.40 ↑1.1	9.30 ↑3.2	31.2	57.3 ↑3.9
+ MemVR	4.80 ↓2.0	3.30 ↓2.0	4.05 ↓2.0	15.0	52.3 ↓1.1
+ Ours	4.80 ↓2.0	3.00 ↓2.3	3.90 ↓2.1	16.6	52.3 ↓1.1

Table 7: Generality Analysis Results. The best results are in green.

(a) Qwen3-VL-8B-Instruct

Method	MME-Perception	LLaVA-Bench
Vanilla	1726	99.5
VCD	1726	101.3
DeCo	1734	77.4
LaSCD	**1740**	**105.8**

(b) MemVR variants

Method	MME-Hall	Perception	Total
Vanilla	643.3	1508.97	1844.68
MemVR (entropy)	648.3	1512.80	1848.51
MemVR + DeCo	633.3	1501.82	1835.03
MemVR + energy	**671.6**	**1533.40**	**1869.47**

Table 8: Ablation results on CHAIR benchmark. The best results are in green.

Methods	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	Average ↓	Len	Recall ↑
LLaVA-1.5	50.0 ↓0.0	15.4 ↓0.0	32.7 ↓0.0	100.6	77.1 ↑0.0
+ Ours	40.6 ↓9.4	11.8 ↓3.6	26.2 ↓6.5	100.0	77.6 ↑0.5
+ LaSCD-α	47.0 ↑6.4	12.3 ↑0.5	29.7 ↑3.5	99.4	78.5 ↑0.9
+ LaSCD-β	43.4 ↑2.8	12.2 ↑0.4	27.8 ↑1.6	98.9	75.3 ↓2.3
+ $\mathcal{K}_\nabla$	43.4 ↑2.8	11.7 ↓0.1	27.6 ↑1.4	98.9	76.8 ↓0.8
+ $\mathcal{K}_{\text{LoG}}$	40.8 ↑0.2	12.0 ↑0.2	26.4 ↑0.2	100.0	77.9 ↑0.3
+ E_{txt}	48.6 ↑8.0	14.2 ↑2.4	31.4 ↑5.2	101.0	81.5 ↑3.9
+ $m^{(l)}$	48.0 ↑7.4	12.9 ↑1.1	30.4 ↑4.2	98.7	78.3 ↑0.7
+ $\text{H}_{\text{Shannon}}$	45.6 ↑5.0	12.1 ↑0.3	28.8 ↑2.6	98.8	78.2 ↑0.6
+ $\mathcal{L} \in \{20\}$	45.4 ↑4.8	12.5 ↑0.7	29.0 ↑2.8	99.1	78.1 ↑0.5
+ $\mathcal{L} \in \{1, \ldots, L\}$	44.0 ↑3.4	12.7 ↑0.9	28.4 ↑2.2	98.3	77.3 ↓0.3

LLaVA- and Qwen-VL–based models, respectively. By contrast, many prior mitigation pipelines sacrifice generality. For example, DeCo causes 89 and 175 point drop on overall perception score for LLaVA-1.5 and GLM-4V. The hallucination performance is improved, yet scores on these broad benchmarks fall. Relative to other CD-based alternatives, LaSCD pairs strong hallucination control with reliable performance on general-purpose evaluation.

5.4 Method Analysis

Ablation Studies. Table 8 verifies that each design choice in LaSCD contributes to hallucination reduction while largely preserving answer coverage. Starting from LLaVA-1.5, LaSCD-α already lowers the average CHAIR score from 32.7 to 29.7 and slightly improves recall, showing that subtracting the hallucination-prone peak-layer preference is an effective first-order correction. Adding the β correction further reduces hallucination, but its lower recall suggests that directly injecting the low-energy layer can become conservative when used alone. Using the Sobel operator $\mathcal{K}_\nabla$ further improves $CHAIR_I$ to 11.7, suggesting that first-order edge responses can also capture hallucination-related attention irregularities; the full Laplacian design gives the best overall average score.

We also compare alternative layer-scoring signals and layer ranges. Using Laplacian of Gaussian kernel $\mathcal{K}_{LoG}$, text-side energy E_{txt}, raw visual mass $m^{(\ell)}$, or Shannon entropy $\text{H}_{\text{Shannon}}$ improves over the baseline, but all are weaker than the full method, supporting our claim that the high-frequency energy is a more reliable hallucination indicator than magnitude or dispersion alone. Likewise, using only a fixed layer $\mathcal{L}=\{20\}$ or all layers $\mathcal{L}=\{1, \ldots, L\}$ underperforms the final design, suggesting that hallucination-related evidence is layer-dependent and benefits from targeted layer selection. Overall, the full LaSCD achieves the best average CHAIR score while keeping response length and recall close to the original model.

Generalization performances. To further demonstrate the generality of the empirical finding, we extend this analysis to more recent and powerful backbone - Qwen3-VL-8B-Instruct. We also decoupled the energy signal from the decoding paradigm and transplanted it into MemVR — an inference-time visual reinforcement paradigm. We changed only MemVR's layer-selection signal from its entropy criterion to our high-frequency energy, keeping all other settings identical. On Qwen3-VL-8B-Instruct, LaSCD improves MME perception and LLaVA-Bench by 14 and 6.3 (Table 7(a)). In the auxiliary MemVR comparison (Table 7(b)), MemVR + DeCo falls below vanilla on MME evaluation, whereas the energy selector improves MME-Hall and total MME by 23.3 and 20.96 over native MemVR (Zou et al., 2025). The energy selector also avoids intermediate LM-head projections, reducing MemVR's per-step complexity.

Table 9: VideoHallucer evaluation results. The best results are in green.

Method	Object–Relation↑	Temporal↑	Semantic↑	Factual↑	Non-Factual↑	Overall↑
Video-LLaVA-7B						
Regular	0.355 ↑0.000	0.097 ↑0.000	0.120 ↑0.000	0.030 ↑0.000	0.260 ↑0.000	0.197 ↑0.000
+ SELF-PEP	0.520 ↑0.165	0.063 ↓0.034	0.360 ↑0.240	0.110 ↑0.080	0.340 ↑0.080	0.432 ↑0.235
+ DeCo	0.355 ↑0.000	0.097 ↑0.000	0.120 ↑0.000	0.030 ↑0.000	0.260 ↑0.000	0.197 ↑0.000
+ Ours	0.410 ↑0.055	0.188 ↑0.091	0.270 ↑0.150	0.150 ↑0.120	0.385 ↑0.125	0.522 ↑0.325
LLaVA-NeXT-Video-7B-DPO						
Regular	0.510 ↑0.000	0.278 ↑0.000	0.380 ↑0.000	0.145 ↑0.000	0.305 ↑0.000	0.609 ↑0.000
+ SELF-PEP	0.520 ↑0.010	0.193 ↓0.085	0.370 ↓0.010	0.115 ↓0.030	0.155 ↓0.150	0.744 ↑0.135
+ DeCo	0.510 ↑0.000	0.193 ↓0.085	0.390 ↑0.010	0.125 ↓0.020	0.285 ↓0.020	0.619 ↑0.010
+ Ours	0.520 ↑0.010	0.193 ↓0.085	0.395 ↑0.015	0.145 ↑0.000	0.300 ↓0.005	0.648 ↑0.039

Table 10: Performance evaluation on MME benchmark and MM-Vet. The best results are in green.

Evaluation	Methods	MME-Hall	Object-Level		Attribute-Level		Perception	Cognition	MM-Vet
		Total ↑	Existence ↑	Count ↑	Position ↑	Color ↑	Total ↑	Total ↑	Accuracy ↑
LLaVA-1.5	Regular	643.3 ↑0.0	190.0 ↑0.0	155.0 ↑0.0	128.3 ↑0.0	170.0 ↑0.0	1508.97 ↑0.0	335.71 ↑0.0	27.50 ↑0.0
	OPERA	610.0 ↓33.	195.0 ↑5.0	128.3 ↓26.	121.7 ↓6.6	165.0 ↓5.0	1473.62 ↓35.	310.71 ↓45.	32.00 ↑4.5
	VCD	648.3 ↑5.0	190.0 ↑0.0	155.0 ↑0.0	133.3 ↑5.0	170.0 ↑0.0	1515.01 ↑6.0	337.86 ↑2.1	26.70 ↓0.8
	ICD	583.3 ↓60.	185.0 ↓5.0	130.0 ↓25.	121.7 ↓6.6	146.7 ↓23.	1306.91 ↓202	287.86 ↓67.	25.90 ↓1.6
	DeCo	608.3 ↓35.	190.0 ↑0.0	135.0 ↓20.	118.3 ↓10.	165.0 ↓5.0	1419.98 ↓89.	304.64 ↓31.	26.10 ↓1.4
	MemVR	648.3 ↑5.0	190.0 ↑0.0	155.0 ↑0.0	133.3 ↑5.0	170.0 ↑0.0	1512.80 ↑9.7	335.71 ↑0.0	32.40 ↑4.9
	AIR	638.3 ↓5.0	190.0 ↑0.0	155.0 ↑0.0	123.3 ↓5.0	170.0 ↑0.0	1504.17 ↓4.8	372.50 ↑36.	-
	DMAS	659.9 ↑16.	195.0 ↑5.0	158.3 ↑3.3	133.3 ↑5.0	173.3 ↑0.0	-	-	-
	Ours	670.0 ↑26.	190.0 ↑0.0	160.0 ↑5.0	145.0 ↑16.	175.0 ↑5.0	1525.92 ↑16.	341.78 ↑6.0	33.40 ↑5.9
Qwen-VL	Regular	600.0 ↑0.0	180.0 ↑0.0	125.0 ↑0.0	130.0 ↑0.0	165.0 ↑0.0	1442.79 ↑0.0	342.14 ↑0.0	31.10 ↑0.0
	OPERA	-	-	-	-	-	-	-	-
	VCD	620.0 ↑20.	175.0 ↓5.0	125.0 ↑0.0	145.0 ↑15.	175.0 ↑10.	1403.17 ↓39.	317.89 ↓24.	24.60 ↓6.5
	ICD	591.6 ↓8.4	170.0 ↓10.	123.3 ↓1.7	138.3 ↑8.3	160.0 ↓5.0	1472.06 ↑29.	351.69 ↑9.5	21.70 ↓9.4
	DeCo	613.3 ↑13.	180.0 ↑0.0	125.0 ↑0.0	133.3 ↑3.3	175.0 ↑10.	1489.53 ↑46.	356.21 ↑14.	31.50 ↑0.4
	MemVR	610.0 ↑10.	180.0 ↑0.0	130.0 ↑5.0	130.0 ↑0.0	170.0 ↑5.0	1473.45 ↑30.	347.50 ↑5.4	33.40 ↑2.3
	AIR	636.7 ↑36.	185.0 ↑0.0	145.0 ↑5.0	126.7 ↑0.0	180.0 ↑5.0	1476.87 ↑34.	352.14 ↑10.	-
	DMAS	633.3 ↑33.	170.0 ↑0.0	145.0 ↑5.0	133.3 ↑0.0	185.5 ↑5.0	-	-	-
	Ours	621.6 ↑21.	180.0 ↑0.0	135.0 ↑10.	141.6 ↑11.	165.0 ↑0.0	1513.79 ↑71.	357.14 ↑15.	32.80 ↑1.7
GLM-4V	Regular	703.3 ↑0.0	200.0 ↑0.0	168.3 ↑0.0	156.7 ↑0.0	178.3 ↑0.0	1680.89 ↑0.0	479.64 ↑0.0	37.60 ↑0.0
	OPERA	-	-	-	-	-	-	-	-
	VCD	675.0 ↓28.	200.0 ↑0.0	166.7 ↓1.6	138.3 ↓18.	170.0 ↓8.3	1624.10 ↓56.	481.43 ↑1.8	33.60 ↓4.0
	ICD	628.3 ↓75.	200.0 ↑0.0	151.7 ↓16.	113.3 ↓43.	163.3 ↓15.	1566.09 ↓114	508.21 ↑28.	32.50 ↓5.1
	DeCo	610.5 ↓92.	200.0 ↑0.0	160.0 ↓8.3	105.5 ↓51.	145.0 ↓33.	1505.86 ↓175	443.57 ↓36.	34.40 ↓3.2
	MemVR	703.3 ↑0.0	200.0 ↑0.0	173.3 ↑5.0	151.7 ↓5.0	178.3 ↑0.0	1683.03 ↑2.1	487.14 ↑7.5	38.20 ↑0.6
	AIR	703.3 ↑0.0	200.0 ↑0.0	173.3 ↑5.0	151.7 ↓5.0	178.3 ↑0.0	1681.89 ↑1.0	479.64 ↑0.0	-
	Ours	709.9 ↑6.6	200.0 ↑0.0	178.3 ↑10.	153.3 ↓3.4	178.3 ↑0.0	1692.02 ↑11.	496.78 ↑17.	38.80 ↑1.2

Inference Latency. LaSCD introduces only a small amount of *additional* compute on top of the *single* autoregressive forward: only two extra decoding head computation to form the contrast decoding. Figure 2 summarizes measured per-token latency, showing that LaSCD remains close to the original model while two-forward CD-style baselines incur statistically higher step time.

6 LIMITATIONS

LaSCD assumes that visual-token order preserves meaningful local structure. This is natural for patch-based encoders with grid-like token layouts, but may be less direct for architectures employing non-grid token aggregation strategies. In addition, our evaluation has not systematically covered reasoning-intensive tasks or substantially larger-scale MLLMs, leaving these settings for future study.

7 CONCLUSION

In this paper, we find that visual hallucination can coexist with high visual-attention mass and that Laplacian energy helps select layers with contrasting answer preferences. Motivated by this finding, we propose LaSCD, a training-free decoding strategy that selects informative layers through high-frequency energy and remaps logits in closed form. Experiments across several benchmarks show hallucination reductions alongside improvements in aggregate general-capability scores. These results motivate further study of visual-attention structure as a signal for hallucination mitigation.

AI USE STATEMENT

This work was completed without the use of any LLMs or AI-assisted tools.

REPRODUCIBILITY STATEMENT

The method and complexity analysis are described in Section 4. Benchmark definitions, inference settings, backbone adaptations, randomization details, computational resources, and artifact information are provided in Appendix B; the complete theoretical argument appears in Appendix A.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

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

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. In *Forty-first International Conference on Machine Learning*, 2024.

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 *The Twelfth International Conference on Learning Representations*, 2023.

Xinpeng Ding, Jianhua Han, Hang Xu, Xiaodan Liang, Wei Zhang, and Xiaomeng Li. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13668–13677, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Team Glm, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. 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*, pp. 14375–14385, 2024.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18135–18143, 2024.

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*, pp. 13418–13427, 2024a.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. Visual hallucinations of multi-modal large language models. *arXiv preprint arXiv:2402.14683*, 2024b.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *ICLR*, 2024.

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*, pp. 13872–13882, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742. PMLR, 2023a.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023b.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023c.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.342. URL https://aclanthology.org/2024.emnlp-main.342/.

Qika Lin, Yifan Zhu, Xin Mei, Ling Huang, Jingying Ma, Kai He, Zhen Peng, Erik Cambria, and Mengling Feng. Has multimodal learning delivered universal intelligence in healthcare? a comprehensive survey. *arXiv preprint arXiv:2408.12880*, 2024b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=J44HfH4JCg.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024c.

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

Yuanhuiyi Lyu, Xu Zheng, Lutao Jiang, Yibo Yan, Xin Zou, Huiyu Zhou, Linfeng Zhang, and Xuming Hu. Realrag: Retrieval-augmented realistic image generation via self-reflective contrastive learning. *arXiv preprint arXiv:2502.00848*, 2025.

Dexter Neo and Tsuhan Chen. Vord: Visual ordinal calibration for mitigating object hallucinations in large vision-language models. *arXiv preprint arXiv:2412.15739*, 2024.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Siya Qi, Yudong Chen, Runcong Zhao, Qinglin Zhu, Zhanghao Hu, Wei Liu, Yulan He, Zheng Yuan, and Lin Gui. Detecting contextual hallucinations in llms with frequency-aware attention. *arXiv preprint arXiv:2602.18145*, 2026.

Xiaoye Qu, Qiyuan Chen, Wei Wei, Jishuo Sun, and Jianfeng Dong. Alleviating hallucination in large vision-language models with active retrieval augmentation. *arXiv preprint arXiv:2408.00555*, 2024.

Lingfeng Ren, Weihao Yu, Runpeng Yu, and Xinchao Wang. Nolan: Mitigating object hallucinations in large vision-language models via dynamic suppression of language priors. *arXiv preprint arXiv:2602.22144*, 2026.

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

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, 2024.

Shezheng Song, Shasha Li, and Jie Yu. Seeing right but saying wrong: Inter-and intra-layer refinement in mllms without training. *arXiv preprint arXiv:2601.07359*, 2026.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. MLLM can see? dynamic correction decoding for hallucination mitigation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4z3IguA4Zg.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024a.

Yujun Wang, Jinhe Bi, Soren Pirk, Yunpu Ma, et al. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 10306–10314, 2026.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arxiv*, 2024b.

Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. Mitigating object hallucination via concentric causal attention. *arXiv preprint arXiv:2410.15926*, 2024.

Xinhao Xu, Hui Chen, Mengyao Lyu, Sicheng Zhao, Yizhe Xiong, Zijia Lin, Jungong Han, and Guiguang Ding. Mitigating hallucinations in multi-modal large language models via image token attention-guided decoding. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *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)*, pp. 1571–1590, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.75. URL https://aclanthology.org/2025.naacl-long.75/.

Diji Yang, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Jie Yang, and Yi Zhang. Im-rag: Multi-round retrieval-augmented generation through learning inner monologues. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 730–740, 2024.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=YtWZdwEG5K.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 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*, pp. 13807–13816, 2024a.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning*, 2024b.

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

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

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

Xingyu Zhu, Kesen Zhao, Liang Yi, Shuo Wang, Zhicai Wang, Beier Zhu, Hanwang Zhang, and Xiangnan He. Look carefully: Adaptive visual reinforcements in multimodal large language models for hallucination mitigation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=KhEjFE1QvQ.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *Forty-second International Conference on Machine Learning (ICML)*, 2025.

A Theoretical Analysis

This section characterizes the Laplacian response under a prescribed neighborhood of visual tokens. We first give its spectral interpretation, then derive a boundary response for region-uniform attention and prove Theorem 1. The analysis holds the neighborhood and attention mass fixed.

A.1 Operator and Spectral Interpretation

Let $\mathcal{V}_g = \{1, \dots, N_v\}$ index the visual tokens. An undirected, unweighted neighborhood is represented by a symmetric adjacency matrix $\mathbf{W}$ with zero diagonal. Write $d_i = \sum_j W_{ij}$, $\mathbf{D} = \mathrm{diag}(d_i)$, and $\mathbf{B} = \mathrm{diag}(b_i)$, where b_i is the nonnegative integer number of connections to fixed zero-valued exterior positions. The operator in Eq. (2) is

$$\Delta_{\mathcal{N}} = \mathbf{D} - \mathbf{W} + \mathbf{B}. \tag{9}$$

Both the neighborhood and $\mathbf{B}$ are fixed independently of the attention values. The boundary term is part of the operator: $\Delta_{\mathcal{N}}\mathbf{1} = \mathbf{B}\mathbf{1}$, so a constant vector can have a nonzero response at the domain boundary.

For normalized attention $\mathbf{p} \geq 0$ with $\mathbf{1}^\top \mathbf{p} = 1$, define

$$Q_{\mathcal{N}}(\mathbf{p}) := \|\Delta_{\mathcal{N}}\mathbf{p}\|_2^2. \tag{10}$$

This squared response is used for the analysis; the decoding score in Eq. (3) averages unsquared per-head response norms.

Spectral representation. The matrix $\Delta_{\mathcal{N}}$ is symmetric positive semidefinite because

$$\mathbf{x}^\top \Delta_{\mathcal{N}} \mathbf{x} = \frac{1}{2} \sum_{i,j} W_{ij}(x_i - x_j)^2 + \sum_i b_i x_i^2 \geq 0. \tag{11}$$

Let $\Delta_{\mathcal{N}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ be an orthonormal eigendecomposition, with eigenvalues $\lambda_k \geq 0$ and eigenvectors $\mathbf{u}_k$. Then

$$Q_{\mathcal{N}}(\mathbf{p}) = \sum_{k=1}^{N_v} \lambda_k^2 \left(\mathbf{u}_k^\top \mathbf{p}\right)^2. \tag{12}$$

Indeed, substituting the eigendecomposition into $\|\Delta_{\mathcal{N}}\mathbf{p}\|_2^2$ and using $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$ gives Eq. (12). Components associated with larger eigenvalues receive greater weight. Thus "high frequency" is defined relative to the chosen neighborhood and boundary convention.

A.2 Boundary Response of Region-Uniform Attention

For any nonempty token subset $R \subseteq \mathcal{V}_g$, consider

$$\mathbf{p}_R = \frac{\mathbf{1}_R}{|R|}. \tag{13}$$

Define the boundary incidence count at each observed token by

$$c_R(i) = \begin{cases} \sum_{j \notin R} W_{ij} + b_i, & i \in R, \\ \sum_{j \in R} W_{ij}, & i \notin R. \end{cases} \tag{14}$$

The exterior positions contribute only through b_i for $i \in R$, since their attention values are fixed to zero. Let $\partial_{\mathcal{N}} R = \{i \in \mathcal{V}_g : c_R(i) > 0\}$ and $d_* = \max_i(d_i + b_i)$.

Proposition (boundary response). For the operator in Eq. (9),

$$Q_{\mathcal{N}}(\mathbf{p}_R) = \frac{\sum_i c_R(i)^2}{|R|^2} \leq \frac{d_*^2 |\partial_{\mathcal{N}} R|}{|R|^2}. \tag{15}$$

Proof. For $i \in R$, contributions from neighbors inside R cancel, leaving $[\Delta_{\mathcal{N}}\mathbf{p}_R]_i = c_R(i)/|R|$. For $i \notin R$, only neighbors inside R contribute, giving $[\Delta_{\mathcal{N}}\mathbf{p}_R]_i = -c_R(i)/|R|$. Squaring and summing proves the equality. Since $c_R(i) \leq d_*$ and it vanishes outside $\partial_{\mathcal{N}} R$, the stated bound follows. □

Comparison at fixed support size. For $|R| = |R'| = s$, Eq. (15) yields the exact comparison

$$Q_{\mathcal{N}}(\mathbf{p}_{R'}) - Q_{\mathcal{N}}(\mathbf{p}_R) = \frac{1}{s^2}\left(\sum_i c_{R'}(i)^2 - \sum_i c_R(i)^2\right). \tag{16}$$

Consequently, a larger squared boundary incidence produces a larger response at fixed support size. Fragmentation increases the response when it increases this boundary quantity, with the effect determined by the connections within the prescribed neighborhood.

A.3 Response to Zero-Mean Perturbations

Let $\boldsymbol{\pi}$ be a fixed normalized attention vector. Consider a random perturbation $\boldsymbol{\eta}$ satisfying

$$\mathbb{E}[\boldsymbol{\eta}] = \mathbf{0}, \qquad \mathbf{1}^\top \boldsymbol{\eta} = 0, \qquad \Sigma_\eta = \mathrm{Cov}(\boldsymbol{\eta}), \tag{17}$$

with finite second moment and $\boldsymbol{\pi} + \boldsymbol{\eta} \geq 0$ almost surely. The sum constraint holds almost surely, so the perturbed vector remains a normalized attention distribution. The operator $\Delta_{\mathcal{N}}$ is held fixed.

Proof of Theorem 1.

Proof. Expanding the squared response gives

$$\begin{aligned}\mathbb{E}\|\Delta_{\mathcal{N}}(\boldsymbol{\pi} + \boldsymbol{\eta})\|_2^2 = \|\Delta_{\mathcal{N}}\boldsymbol{\pi}\|_2^2 \\ + 2(\Delta_{\mathcal{N}}\boldsymbol{\pi})^\top \Delta_{\mathcal{N}}\mathbb{E}[\boldsymbol{\eta}] + \mathbb{E}[\boldsymbol{\eta}^\top \Delta_{\mathcal{N}}^\top \Delta_{\mathcal{N}}\boldsymbol{\eta}].\end{aligned} \tag{18}$$

The cross term is zero. Since $\mathbb{E}[\boldsymbol{\eta}\boldsymbol{\eta}^\top] = \Sigma_\eta$, the remaining term is

$$\mathbb{E}[\boldsymbol{\eta}^\top \Delta_{\mathcal{N}}^\top \Delta_{\mathcal{N}}\boldsymbol{\eta}] = \mathrm{Tr}(\Delta_{\mathcal{N}}^\top \Delta_{\mathcal{N}}\Sigma_\eta) \geq 0. \tag{19}$$

It is strictly positive precisely when $\mathbb{E}\|\Delta_{\mathcal{N}}\boldsymbol{\eta}\|_2^2 > 0$. This proves the identity and strict-increase condition in Eq. (8). □

For the unsquared response, convexity of the norm also gives

$$\mathbb{E}\|\Delta_{\mathcal{N}}(\boldsymbol{\pi} + \boldsymbol{\eta})\|_2 \geq \|\Delta_{\mathcal{N}}\boldsymbol{\pi}\|_2. \tag{20}$$

This norm inequality and the squared-response identity describe expectations over perturbations; individual realizations can increase or decrease the response.

A.4 Relation to the Decoding Score

The normalized analysis isolates local variation. For a nonzero raw visual slice, write $\boldsymbol{\alpha}_{\mathrm{vis}}^{(\ell,h)} = m^{(\ell,h)}\mathbf{p}^{(\ell,h)}$, where $m^{(\ell,h)} = \|\boldsymbol{\alpha}_{\mathrm{vis}}^{(\ell,h)}\|_1$. By linearity, Eq. (3) is exactly

$$E^{(\ell)} = \frac{1}{H}\sum_{h=1}^{H} m^{(\ell,h)}\sqrt{Q_{\mathcal{N}}(\mathbf{p}^{(\ell,h)})}. \tag{21}$$

Zero-mass slices contribute zero. Thus the implemented score is the head average of mass-weighted response norms, relating the normalized analysis to attention with varying visual mass.

These results characterize the operator's spectral weighting, boundary sensitivity, and response to stochastic variation under one neighborhood-based definition. The response depends on attention geometry: small or multiple relevant regions can also yield high energy, while smooth attention can cover irrelevant regions. The empirical analysis in Sec. 3.3 examines how this response relates to layer-wise answer preferences.

B Experimental Details

B.1 Benchmarks and Metrics

Hallucination benchmarks. We evaluate image-level hallucination on POPE (Li et al., 2023c), CHAIR (Rohrbach et al., 2018), and HallusionBench (Guan et al., 2024), and further evaluate video hallucination on VideoHallucer (Wang et al., 2024b). POPE is a binary VQA-style benchmark that probes object existence with questions of the form "Is [object] in this image?"; we append the instruction "Please answer yes or no." and report performance under the random, popular, and adversarial splits. CHAIR evaluates object hallucination in generated captions. Following standard practice, we randomly select 500 images from the MSCOCO Val2014 split (Lin et al., 2014) and the prompt "Please describe this image in detail." We report CHAIR_S, which measures the fraction of captions containing at least one hallucinated object, and CHAIR_I, which measures the fraction of mentioned objects that are hallucinated:

$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}, \qquad \text{CHAIR}_S = \frac{|\{\text{sentences with hallucinated objects}\}|}{|\{\text{all sentences}\}|}. \tag{22}$$

HallusionBench evaluates multimodal hallucination with a broader set of visual reasoning and language biases; we follow the official evaluation protocol. VideoHallucer evaluates hallucination in large video-language models with adversarial binary VideoQA pairs, covering both intrinsic hallucinations, such as object-relation and temporal hallucinations, and extrinsic hallucinations, including factual and non-factual hallucinations.

General multimodal benchmarks. To assess the impact of hallucination mitigation on general ability, we evaluate on MME (Fu et al., 2023), MM-Vet (Yu et al., 2024b), and LLaVA-Bench-in-the-wild (Liu et al., 2024c). MME contains perception and cognition categories and uses yes/no questions. MM-Vet evaluates integrated multimodal capabilities with open-ended responses. For LLaVA-Bench-in-the-wild, we follow the GPT-based evaluation protocol used by the benchmark. All benchmarks that require external model-based judging, including MM-Vet, HallusionBench and LLaVA-Bench-in-the-wild, are evaluated with the GPT-4 API.

B.2 Inference Protocol

We use greedy decoding as the default decoding strategy for LaSCD and the original MLLM unless a baseline requires sampling or beam search by design. Questions from each benchmark are formatted using the chat template of the evaluated MLLM. For binary VQA-style benchmarks such as POPE, MME, and VideoHallucer, we parse the model output into yes/no answers; for caption-style evaluation such as CHAIR, we use the standard object-matching protocol. For GPT-based evaluations, we use the official or benchmark-recommended evaluator when available and run all external model-based judging with the GPT-4 API.

Baseline settings. We compare against training-free hallucination mitigation methods that intervene at decoding time. Our baseline reproduction follows MemVR (Zou et al., 2025) in evaluation protocol and practice: whenever a method releases official code, we re-run it with its released scripts and paper-aligned hyper-parameters; when no implementation is available, we copy the numbers directly from the original paper's tables. For any model–benchmark pair that a baseline paper does not report, we leave the corresponding table entry blank rather than extrapolating. For DeCo (Wang et al., 2025) specifically, we *manually reproduce* every dataset–model combination absent from its published evaluations. For the auxiliary check in Table 7(b), we retain MemVR's visual-reinforcement module and inference settings and replace only its entropy-based layer-selection criterion with $E^{(\ell)}$.

LaSCD settings. For LaSCD and the original MLLM, we use greedy decoding by default and keep the same prompt template, image/video preprocessing, maximum generation length, and answer parsing within each benchmark. During LaSCD decoding, the model is run once per generation step with attentions and hidden states returned, so that we can read both the final prediction and the intermediate-layer predictions without a second forward pass.

The algorithm has three parameterized components. First, for **layer selection**, we apply $\Delta_{\mathcal{N}}$ to the raw visual-attention slice of each candidate layer and average the per-head response norms, as in Eq. (3). The neighborhood and zero-extension boundary convention are specified in Appendix B.4.

For all the MLLMs, the candidate layer range is set within $[8, 29)$. Alternative neighborhoods, Sobel and Laplacian-of-Gaussian operators, and text-token energy are used only for ablation.

Second, for **candidate-token filtering**, we restrict logit remapping to a plausible token set selected from the final-layer distribution. Unless otherwise stated, we use top-k=10 and cumulative probability threshold p=0.9. Remapping operates within the candidate set retained by these thresholds.

Third, for **logit remapping**, α controls the strength of suppressing the high-energy layer's preference. We set $\alpha = 0.1$ in the main experiments. We set $\beta = 0.2$ for HallusionBench and 0.6 for CHAIR evaluation. All parameters are fixed for full evaluation.

B.3 Backbone Adaptation and Compatibility

LaSCD requires two kinds of intermediate signals from the backbone: the last-query attention over visual tokens and the LM-head logits from selected intermediate layers. Different MLLM implementations expose these signals differently, so we use lightweight model-specific adapters while keeping the original image/video encoder, prompt template, and tokenizer unchanged.

LLaVA-Series. For LLaVA-1.5, LLaVA-Next-Video-DPO and Video-LLaVA, we insert the same image/video placeholder tokens used by the original model and force the attention implementation to an eager mode so that attention maps can be captured during generation. The adapter records the expanded multimodal sequence length after visual features are inserted and uses the placeholder position to align the visual-token span in the KV sequence. For video models, the visual tokens are produced from frame-level features and remain contiguous in the multimodal sequence; Video-LLaVA explicitly uses one image placeholder per sampled frame, while LLaVA-Next-Video uses the model's multimodal preparation routine to expand the video input. Intermediate-layer logits are obtained from the same forward pass through a forward-capture hook, avoiding an additional decoding pass.

GLM-4V. GLM-4V does not expose attention in the same format as LLaVA-style models. We therefore keep its original core-attention forward path unchanged and recompute the attention probabilities in a side channel for the monitored layers. This design avoids changing the hidden-state trajectory when the intervention strength is zero. The visual-token span is located using the begin-of-image marker when available; otherwise, we infer the number of inserted visual tokens from the KV length after multimodal expansion and use a conservative fallback start position. Intermediate logits are obtained by applying the model's output layer to hooked block hidden states, with the final layer normalization used when required by the backbone.

Qwen-VL. For Qwen-VL, we use the model's returned self-attention tensors and register hooks on the selected transformer blocks to obtain intermediate hidden states. The visual-token span is identified from the tokenizer's image start/end markers, matching the model's convention that image embeddings replace the tokens between these markers. Intermediate logits are computed by applying the final normalization and output embedding head to the hooked hidden states.

B.4 Neighborhood Construction and Boundary Handling

The reported configuration uses unit-weight local neighbors determined by visual-token indexing. Square-length visual spans are arranged in row-major order, with immediate horizontal and vertical positions defining $\mathcal{N}(i)$ and $b_i = 4 - |\mathcal{N}(i)|$. Other spans use predecessor and successor positions in the visual-token sequence, with $b_i = 2 - |\mathcal{N}(i)|$. Missing positions have fixed zero values. Both constructions instantiate Eq. (2); they use the same response norm, head aggregation, and layer-selection rule.

The sequence-neighbor ablation, denoted $\mathcal{N}_{\text{seq}}$ in Table 8, uses predecessor and successor positions for every visual span. Its score is still Eq. (3), with $\Delta_{\mathcal{N}_{\text{seq}}}$ as the operator. The sign convention in Eq. (2) has a positive central coefficient; reversing the sign of the entire stencil leaves the response norm unchanged. Vectorizing a convolution response also preserves its norm, since the Frobenius norm of an array equals the Euclidean norm of its vectorization.

These neighborhoods are defined by token indexing. Their interpretation as image-space locality depends on the backbone: raster-ordered patch tokens preserve local spatial relations, whereas resampling, padding removal, or frame concatenation can alter them. The score measures local variation under the resulting indexing convention.

B.5 Reproducibility Details

We fix random seeds for sampling, subset selection, and repeated evaluation when applicable. For CHAIR, a randomly sampled image set from the COCO Val2014 dataset was used across all evaluations. We sampled five different sets of images using different random seeds and evaluated performance by calculating the average results. For POPE, we evaluate all three sampling protocols (random, popular, and adversarial). For MME and VideoHallucer, we use yes/no answer parsing and skip non-binary samples only in analyses where binary correctness is required. For reproduced baselines, we match image or video inputs, prompts, and answer parsing rules within each benchmark.

For analysis figures that inspect layer-wise behavior, we run the model with both hidden states and attentions returned. We compute per-layer logits by applying the LM head to normalized intermediate hidden states and compute high-frequency energy on the last-query visual-attention slice. These analysis runs are used only to diagnose layer behavior and produce figures; main benchmark evaluation uses the decoding protocol described above.

B.6 Computational Resources

All experiments are implemented in PyTorch (Paszke et al., 2019) and conducted on NVIDIA A800 80GB GPUs.

B.7 Artifacts and Licenses

We follow the licenses of all datasets, model checkpoints, and evaluation tools used in the experiments, and restrict their use to academic research when required. The evaluated datasets and benchmarks include MSCOCO, POPE, CHAIR, HallusionBench, VideoHallucer, MME, MM-Vet, and LLaVA-Bench-in-the-wild. The evaluated model families include LLaVA, Qwen-VL, Qwen3-VL, GLM-4V, InstructBLIP, Video-LLaVA, and LLaVA-Next-Video, subject to the licenses of their released checkpoints and code. We will release code and instructions for reproducing the reported evaluations and analysis figures.

C Limitations

LaSCD relies on intermediate attention signals, which creates several practical limitations. First, Laplacian energy measures local variation relative to a prescribed token neighborhood and boundary convention. Its interpretation depends on whether that neighborhood preserves useful relations between visual tokens. Aggressive token pooling, learned resamplers, token pruning, or mixing across frames may weaken the correspondence between index-based neighbors and regions in the visual input. Sensitivity to these transformations motivates studying neighborhood construction across backbones. Second, implementation requires access to self-attention maps and intermediate-layer logits, potentially requiring adapters in serving systems. Third, the effect of the intervention on reasoning-intensive tasks warrants systematic evaluation. Finally, extending the current image and video evaluations to substantially larger-scale MLLMs would help characterize how the method behaves as model capacity increases.

Failure Cases. We emphasize that high-frequency Laplacian energy is not an absolute marker of token-level hallucination: in our offline inspection we did observe failure cases in both directions, where some high-energy tokens are in fact grounded and some low-energy tokens are in fact hallucinated. Although high-frequency energy show a relatively strong statistical correlation with hallucinatory behavior, they cannot be regarded as the root cause of visual hallucinations in multimodal models. However, LaSCD's design with its contrastive-decoding formulation, ensure that such residual failures do not compromise inference performance.

D Case Study

In this section, we present additional cases illustrating the layer-wise attention and answer preferences examined in Sec. 3.2 and Sec. 3.3.

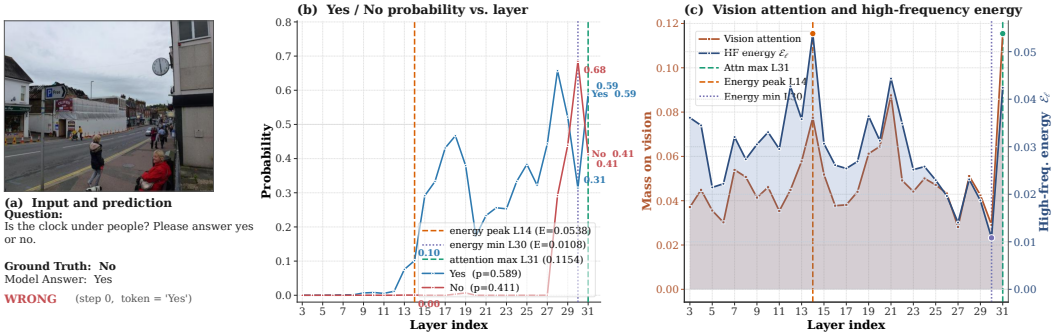

Figure 5: A case study of hallucination during decoding progress.

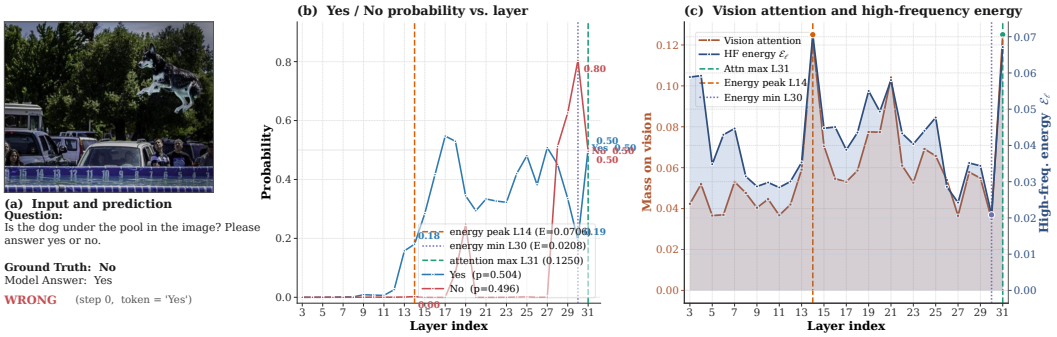

Figure 6: A case study of hallucination during decoding progress.

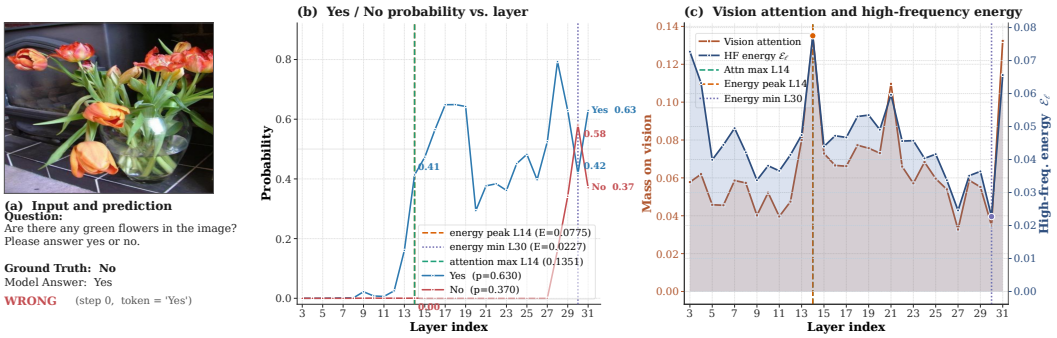

Figure 7: A case study of hallucination during decoding progress.

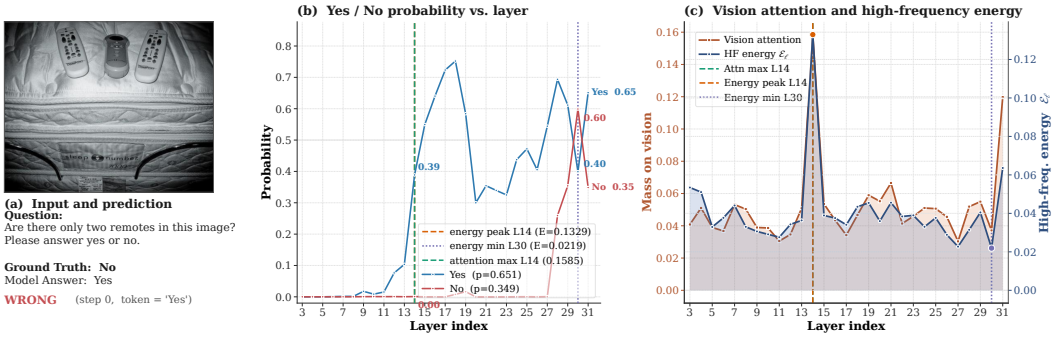

Figure 8: A case study of hallucination during decoding progress.

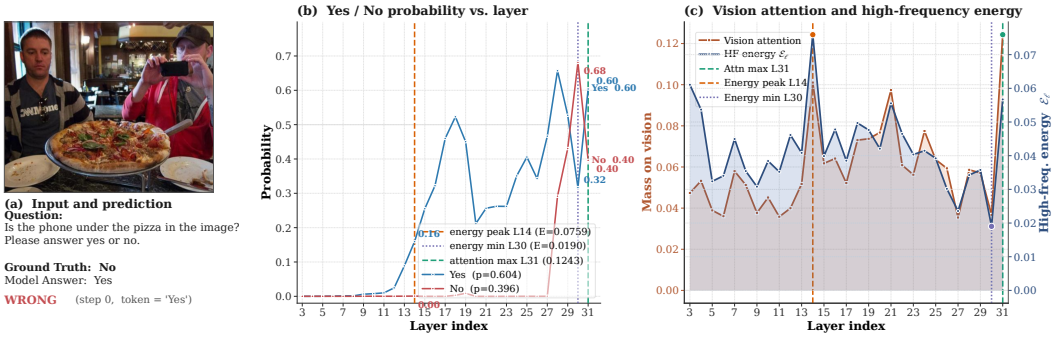

Figure 9: A case study of hallucination during decoding progress.

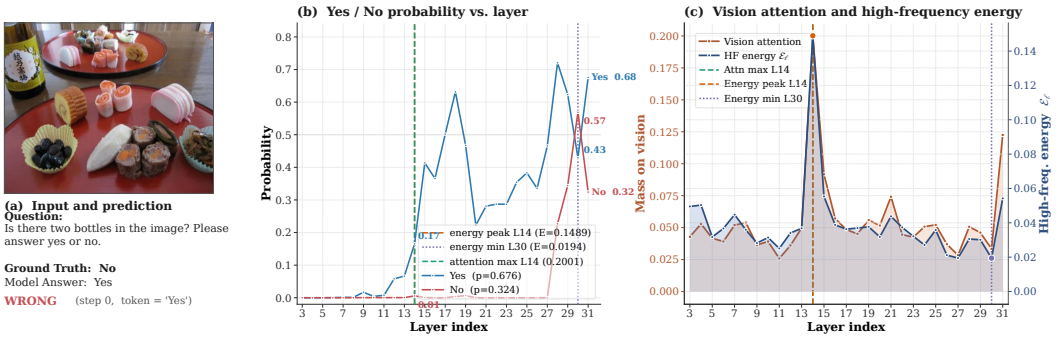

Figure 10: A case study of hallucination during decoding progress.

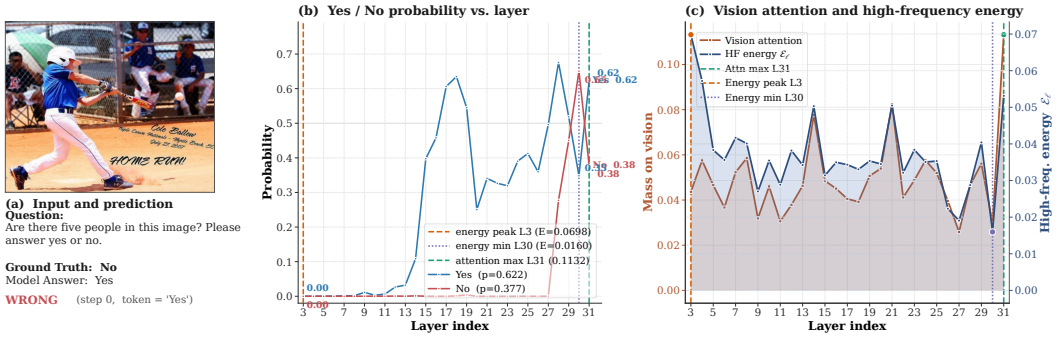

Figure 11: A case study of hallucination during decoding progress.