HSS-DR: FROM HIDDEN-STATE SPECTRAL SIGNATURES TO OBJECT HALLUCINATION DETECTION AND REPAIR IN LVLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (LVLMs) have demonstrated strong capabilities in multimodal understanding, yet they still hallucinate objects that are absent from the input image. We investigate this problem by analyzing the spatial spectra of visual representations computed before response generation. In late decoder layers, these spectra show a lower proportion of low-frequency energy and a higher spectral centroid for subsequent hallucinated mentions than for grounded mentions. Based on this finding, we propose HSS-DR, a framework for object hallucination detection and repair. We train a lightweight detector that uses spectral features to identify object mentions at risk of hallucination. For each flagged object mention, HSS-DR uses constrained search to find a lower-risk continuation. The same detector evaluates candidate continuations over a short lookahead and guides their selection based on hallucination risk. Experiments across multiple LVLMs demonstrate that HSS-DR consistently reduces object hallucination while preserving general multimodal capabilities.

1 INTRODUCTION

Large vision-language models (LVLMs) have demonstrated strong capabilities in image understanding and language generation, supporting tasks such as image captioning and visual question answering (Li et al., 2023a; Liu et al., 2023). However, they still suffer from object hallucination, generating descriptions of objects that are absent from the input image (Rohrbach et al., 2018; Li et al., 2023b). Such errors undermine the reliability of their responses and can mislead users (Ho et al., 2026). Investigating the internal representations associated with object hallucination is therefore important for developing effective mitigation methods.

Existing approaches mitigate object hallucination through interventions on decoding or internal representations. Contrastive decoding adjusts output distributions using perturbed visual inputs (Leng et al., 2024), while attention-based methods strengthen the model's use of image content (Liu et al., 2025c). Interventions in latent space further exploit hidden representations to suppress components associated with hallucination or reinforce visual evidence (Liu et al., 2025b; Duan et al., 2025). These methods demonstrate the value of internal signals for improving visual grounding. We investigate whether visual hidden representations encode spatial patterns that distinguish grounded objects from hallucinated ones. Identifying such patterns would provide a signal for detecting unreliable object mentions and guiding local intervention during generation.

We hypothesize that grounded objects are associated with more spatially coherent responses across supporting image regions, whereas hallucinated objects are associated with more irregular patterns. Spatial spectral analysis characterizes these patterns through their variation across image locations: low-frequency components capture slowly varying response structure, while higher-frequency components capture more rapid spatial variations. We therefore examine how grounded and hallucinated objects distribute their response energy across spatial frequencies.

We construct object-direction response maps from visual hidden states available before response generation. At each decoder layer, we project visual-token states onto an object direction and arrange the responses on the image patch grid. Our analysis of LLaVA-1.5 reveals that these visual representations already exhibit different spatial spectra for grounded and hallucinated mentions in

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.

the subsequent response. Hallucinated mentions are associated with a lower proportion of low-frequency energy and a higher spectral centroid than grounded mentions. Spectral differences are weak in shallow decoder layers and become pronounced in late layers. The same tendency appears across our analysis splits and in comparisons between grounded and hallucinated objects within the same image. This spectral pattern also appears in the late decoder layers of Qwen-VL-Chat, extending the evidence to a different visual-token architecture (Appendix C.4). These observations identify spectral signatures of object hallucination in hidden states and provide an empirical basis for estimating object-level risk.

Building on these findings, we propose HSS-DR, a framework for object hallucination detection and local repair. We train a lightweight detector that uses spectral features, complemented by response-map features and generation context, to estimate hallucination risk for individual object mentions. When a mention is flagged, HSS-DR searches for a lower-risk continuation under language model constraints. The same detector evaluates candidate continuations over a short lookahead to guide candidate selection. The LVLM backbone remains frozen, with training confined to the lightweight detector.

We evaluate HSS-DR across multiple LVLMs with different architectures. Across all six models, HSS-DR achieves a 24.2%–37.0% relative reduction in the sentence-level hallucination rate (CHAIR_S) compared with vanilla decoding. Evaluations on AMBER further demonstrate its effectiveness in open-ended generation, while results on MM-Vet show that HSS-DR preserves general multimodal capabilities. Ablation studies examine the roles of detector inputs and repair components.

Our contributions are threefold:

- We identify hidden-state spectral signatures that distinguish grounded from hallucinated object responses in late decoder layers, with consistent patterns across different visual-token architectures.
- We develop HSS-DR, which uses a lightweight detector informed by these spectral signatures to identify high-risk object mentions. The same detector guides local repair by evaluating candidate continuations over a short lookahead.
- We demonstrate consistent reductions in object hallucination across six LVLMs while preserving general multimodal capabilities.

2 Related Work

2.1 Internal Representations and Object Hallucination

Research on internal representations addresses two related questions: how LVLMs use visual evidence, and whether their internal states reveal hallucination. Layerwise analyses track the evolution of visual information and object predictions (Neo et al., 2025; Wang et al., 2025), while attention-based studies examine how information is allocated and aggregated during generation (Jiang et al., 2025; Kang et al., 2025; Tang et al., 2025; Yang et al., 2025). Complementary analyses study the stability of visual features (Liu et al., 2025b), the influence of preceding generated tokens (Zhang et al., 2026b), and the separation of visual evidence from language priors (Lin et al., 2026; Wu et al., 2026). Learned detectors further use token representations (Duan et al., 2025; Park et al., 2025), attention-head outputs (Zhang et al., 2026a), or neuron activations (Kim et al., 2025) to identify content without sufficient visual support.

MFCD applies frequency filtering to input images and contrasts the resulting output distributions with that of the original image (Liu et al., 2025a). We instead examine spatial spectra of object-related responses in visual hidden states, using differences between grounded and hallucinated objects as a signal for object-level risk estimation.

2.2 Object Hallucination Mitigation

Existing methods improve visual grounding through training, inference-time intervention, or response revision. Training-based approaches use additional feedback (Chen et al., 2024a) or preference alignment (Sun et al., 2024; Yu et al., 2024a; He et al., 2025). Contrastive decoding (Leng et al.,

2024; Wang et al., 2024b; Huo et al., 2025) and prediction ensembling (An et al., 2025) recalibrate output distributions during inference. Other methods exploit intermediate-layer information (Wang et al., 2025; Li et al., 2025), adjust visual attention (Liu et al., 2025c; Kang et al., 2025), or modify latent activations (Liu et al., 2025b; Zou et al., 2025). After generation, learned revisers (Zhou et al., 2024) and external visual verification (Yin et al., 2024) correct unsupported content in completed responses.

Hallucination signals also guide corrections during generation. OPERA combines an attention-based penalty with rollback and token reallocation (Huang et al., 2024). REVERSE trains the VLM to verify generated phrases and uses the resulting confidence signal to trigger retrospective resampling (Wu et al., 2025). The Visual Absence detector guides rollback and token replacement using feed-forward network activations (Kim et al., 2025). HSS-DR brings spatial spectral features into this setting: the same lightweight detector identifies high-risk object mentions and evaluates candidate continuations over a short lookahead to guide local repair.

3 Spatial Spectral Signatures of Object Hallucination

We investigate how the spatial organization of visual hidden representations differs between grounded and hallucinated objects. Our analysis uses LLaVA-1.5-7B (Liu et al., 2024) and follows visual representations across its 32 decoder layers.

3.1 From Visual Hidden States to Spatial Spectra

We analyze 739 object mentions generated for 100 COCO (Lin et al., 2014) development images: 636 grounded and 103 hallucinated. Mentions are labeled by whether the corresponding object is present in the image. For each image, we extract visual hidden states from the initial forward pass, using the prompt "Please describe this image in detail." before response generation. The same states are used to analyze all object mentions in the subsequent description.

Let $\widetilde{\mathbf{h}}_{u,v}^{(\ell)}$ denote the visual hidden state at patch location (u, v) and decoder layer ℓ, expressed through the model's final normalization. For an object surface form a with tokens $(t_1, \ldots, t_m)$, we construct its direction from the corresponding rows of the language model output matrix W:

$$\mathbf{d}_a = \mathrm{normalize}\left(\frac{1}{m}\sum_{j=1}^{m}\frac{W_{t_j}}{\|W_{t_j}\|_2}\right), \qquad M_a^{(\ell)}(u,v) = \left\langle \widetilde{\mathbf{h}}_{u,v}^{(\ell)}, \mathbf{d}_a \right\rangle. \tag{1}$$

Here, $\mathrm{normalize}$ denotes ℓ_2 normalization. The resulting *object-direction response map* $M_a^{(\ell)}$ retains the spatial arrangement of the visual tokens. We restrict the original 24×24 patch grid to its image-content region, excluding padding. Each map describes how the hidden representation responds to a particular object across image locations.

We apply an orthonormal two-dimensional DCT-II (Ahmed et al., 1974) and normalize its squared coefficients by their total power:

$$C_a^{(\ell)} = \mathrm{DCT}_2\left(M_a^{(\ell)}\right), \qquad P_a^{(\ell)}(k,q) = \frac{\left|C_a^{(\ell)}(k,q)\right|^2}{\max\left(\sum_{k',q'}\left|C_a^{(\ell)}(k',q')\right|^2, \epsilon\right)}, \tag{2}$$

where $\epsilon = 10^{-8}$. The transform preserves the DC component. We partition DCT modes into six radial bands $B_0, \ldots, B_5$ containing approximately equal numbers of modes, ordered from low to high frequency. Writing $p_{a,b}^{(\ell)} = \sum_{(k,q)\in B_b} P_a^{(\ell)}(k,q)$, we summarize each spectrum by its low-frequency energy proportion ρ and spectral centroid μ:

$$\rho_a^{(\ell)} = p_{a,0}^{(\ell)} + p_{a,1}^{(\ell)}, \qquad \mu_a^{(\ell)} = \sum_{b=0}^{5}\frac{b}{5}\,p_{a,b}^{(\ell)}. \tag{3}$$

Low frequencies describe slowly varying spatial responses, while higher frequencies capture more rapid variations. Figure 1 illustrates these components for a grounded *car* and a hallucinated *traffic light* in the same image. The car response contains a coherent region retained by the low-frequency reconstruction, whereas the traffic-light response is more irregular.

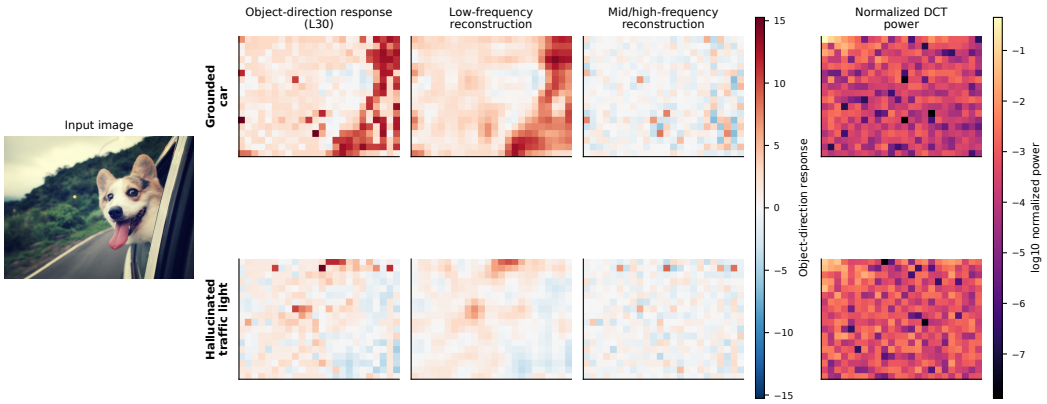


Figure 1: **Object-level spatial and spectral responses at L30.** A development-split image is evaluated along a grounded car direction (top) and a hallucinated traffic-light direction (bottom). Columns show the response map, reconstructions using B_0–B_1 and B_2–B_5, and normalized DCT power on a logarithmic scale.

3.2 Spectral Separation across Decoder Layers

Figure 2 compares the mean band energies of the two classes at every decoder layer. Differences are small in shallow layers and become pronounced in intermediate-to-late layers. In the later layers, hallucinated objects have less energy in the lowest band B_0, accompanied by higher proportions in B_1–B_5. Although B_1 gains energy, the larger reduction in B_0 yields a lower combined low-frequency proportion.

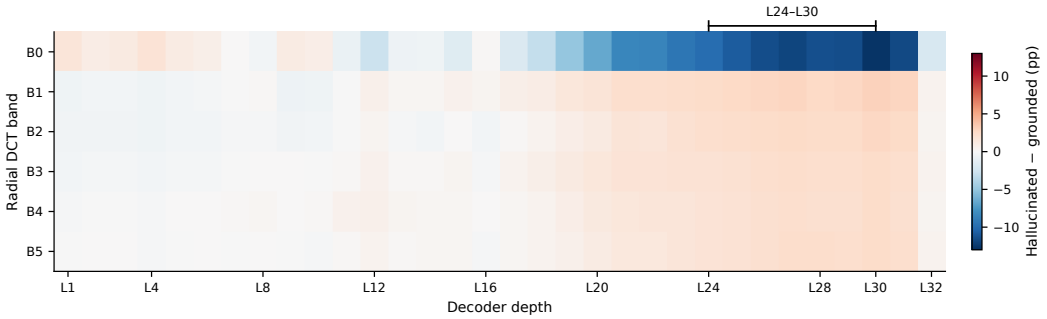


Figure 2: **Layer-wise spectral energy differences.** Hallucinated-minus-grounded mean energy in each radial DCT band on the development split, in percentage points. The bracket marks L24–L30.

The class trajectories in Figure 3 clarify how this separation develops. Grounded responses become more concentrated at low frequencies as depth increases, while hallucinated responses retain relatively more energy at higher frequencies. Averaged over L24–L30, grounded mentions have an 8.76 percentage point higher low-frequency proportion, and hallucinated mentions have a 0.0658 higher spectral centroid. Both class gaps peak at L30 and contract at the final layer. The pronounced separation across L24–L30 provides an empirical basis for extracting response-map features from this window in HSS-DR.

The spectral separation is robust across data partitions. The same trend is reproduced on a disjoint confirmation split, and each spectral statistic follows the expected ordering in over 84% of within-image comparisons containing both grounded and hallucinated objects (Appendices C.1 and C.2).

These findings motivate a lightweight detector that combines features derived from object-specific response maps with generation context to estimate hallucination risk and guide local repair.

4 METHOD

HSS-DR uses the spatial spectral signatures identified in Section 3 to detect and repair object hallucinations during generation. A lightweight detector combines features derived from object-direction response maps with generation context to estimate the risk of an object mention. This estimate determines when to revisit the mention and how to select a local continuation. We describe the method using LLaVA-1.5, whose parameters remain frozen throughout detector training and inference. Appendix A.1 describes adaptations to other architectures. Figure 4 summarizes the framework.

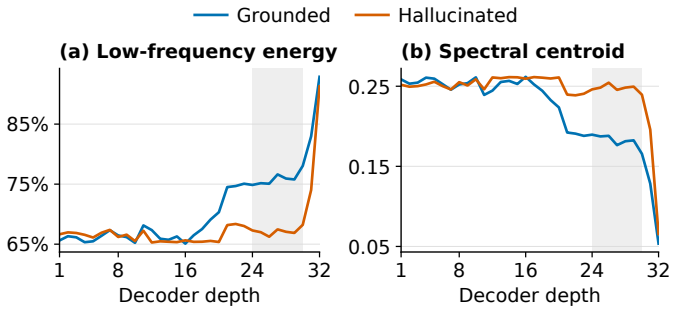


Figure 3: **Class trajectories across decoder layers.** Means use the development split; shading marks L24–L30.

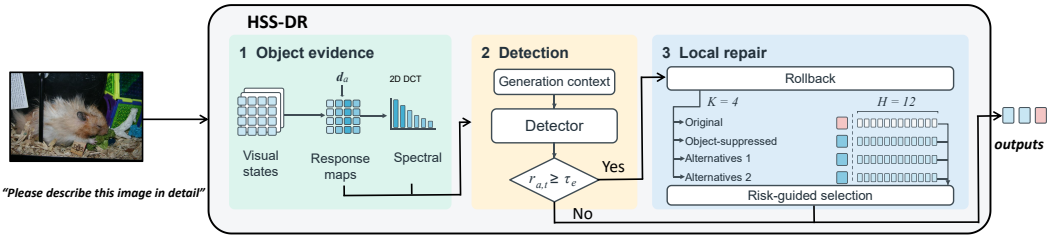


Figure 4: **Overview of HSS-DR.** Spectral and response-map features are combined with generation context to flag high-risk object mentions. Local repair compares $K = 4$ candidate continuations, each scored by the same detector over $H = 12$ post-mention tokens. Selection favors lower hallucination risk while limiting deviation from the model's token preferences. Generation resumes from the selected continuation. Grids, spectra, and token sequences are schematic.

4.1 OBJECT HALLUCINATION DETECTION

Representing object evidence. Motivated by the separation observed in Section 3.2, we extract response features from $\mathcal{L} = \{24, \ldots, 30\}$. For each object surface form a, the feature vector $\mathbf{z}_a$ summarizes the frequency distribution, magnitude, and spatial concentration of its responses, together with their variation across layers and consistency across forms of the same category. These features are computed from the response maps and spectra defined in Section 3.1, then cached for reuse throughout generation (Appendix B.1).

Estimating mention risk. During greedy decoding, we score an object candidate when the model's preferred token completes a surface form in the object vocabulary. At completion position t, we combine its cached representation with generation context $\mathbf{c}_{a,t}$, which summarizes decoding probabilities and position:

$$\mathbf{x}_{a,t} = [\mathbf{z}_a; \mathbf{c}_{a,t}], \qquad r_{a,t} = G_e(f_{\theta_e}(\mathbf{x}_{a,t})) \,. \tag{4}$$

Here, e selects one of two lightweight gradient-boosted tree classifiers for single-token and multi-token surface forms. We train them on generated object mentions, using object presence in the image as supervision. The mapping G_e converts the classifier score to a percentile relative to grounded training mentions, with higher values indicating greater hallucination risk.

Calibrating intervention. We calibrate an expert-specific threshold τ_e for high recall on a held-out split. Intervention is deferred until the next predicted token confirms the mention's word boundary; $r_{a,t} \geq \tau_e$ then triggers local repair, subject to the search budget. Training and calibration details are provided in Appendix B.1.

4.2 Risk-Guided Local Repair

Constructing local continuations. When a mention triggers repair, we roll back to the prefix preceding its first token. We retain the original greedy continuation as the identity candidate. A second candidate uses greedy decoding while blocking tokens that would complete a mention of the flagged object category. Two further candidates start from distinct high-probability alternatives to the second candidate's initial token and continue greedily under the same constraint, giving $K = 4$ candidates in total. All candidates share an evaluation endpoint $H = 12$ tokens beyond the original mention's completion position. Within this horizon, the same detector evaluates object mentions using their cached spectral and response-map features together with decoding probabilities and positions along each candidate. These scores measure the hallucination risk of individual object mentions and guide the comparison of candidate continuations.

Ranking candidate continuations. For each candidate continuation b, we assess hallucination risk using the number of unsafe object mentions and their excess risk above the calibrated threshold. Let $R_{\text{hall}}(b)$ denote this ordered risk summary and $R_{\text{LM}}(b)$ the mean language model regret, measured as the log probability gap between unconstrained greedy choices and selected tokens along b. The eligible set $\mathcal{B}$ contains the identity continuation and valid repair candidates with bounded regret. We select

$$b^* = \arg\min_{b \in \mathcal{B}}^{\text{lex}} \left(R_{\text{hall}}(b), R_{\text{LM}}(b) \right), \tag{5}$$

where lexicographic ordering prioritizes lower hallucination risk, followed by smaller deviation from the model's token preferences.

Accepting and committing a repair. The ranking identifies the preferred candidate. We commit a repair only if it achieves the required risk reduction relative to the identity continuation; otherwise, we retain the identity continuation. Greedy decoding then resumes from the accepted continuation, subject to the remaining search budget.

5 Experiments

5.1 Experimental Setup

Models and baselines. We evaluate HSS-DR on six LVLMs: LLaVA-1.5-7B (Liu et al., 2024), Qwen-VL-Chat (Bai et al., 2023), MiniGPT-4 with Vicuna-7B (Zhu et al., 2024), InstructBLIP with Vicuna-7B (Dai et al., 2023), Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), and Qwen3-VL-8B-Instruct (Bai et al., 2025a). Each backbone has its own trained detector. We compare against vanilla decoding and five hallucination mitigation methods: VCD (Leng et al., 2024), DeCo (Wang et al., 2025), FLB (Ha et al., 2026), SHIELD (Huang et al., 2026), and MFCD (Liu et al., 2025a). The result tables specify the evaluated models and baselines for each benchmark.

Benchmarks and metrics. We assess object hallucination and coverage using CHAIR, OPOPE, and AMBER. For CHAIR, we evaluate detailed image descriptions on a fixed set of 500 COCO (Lin et al., 2014) images. CHAIR_S measures the fraction of descriptions containing hallucinated objects, and CHAIR_I measures the fraction of object mentions that are hallucinated (Rohrbach et al., 2018). We also report object Recall to measure coverage of ground-truth objects. OPOPE (Chen et al., 2024b) evaluates the same descriptions by checking whether queried objects appear in their extracted object sets. We report Precision, Recall, and $F_{0.2}$ under the Random, Popular, and Adversarial sampling schemes of POPE (Li et al., 2023b); $F_{0.2}$ gives greater weight to precision. AMBER Generative (Wang et al., 2024a) assesses hallucination and object coverage on 1,004 responses per method using CHAIR, Cover, Hal, and Cog.

Table 1: **Object hallucination on CHAIR and OPOPE.** All scores are percentages. C_S and C_I denote CHAIR_S and CHAIR_I; P denotes Precision and R denotes Recall. Recall is computed separately for the two benchmarks and is shared across the three OPOPE splits. Bold marks the best displayed score within each model.

Model	Method	CHAIR			OPOPE						
					Random		Popular		Adversarial		R↑
		$C_S\downarrow$	$C_I\downarrow$	R↑	P↑	$F_{0.2}\uparrow$	P↑	$F_{0.2}\uparrow$	P↑	$F_{0.2}\uparrow$	
LLaVA-1.5	Vanilla	51.4	14.40	**81.47**	98.41	97.74	90.73	90.43	87.99	87.81	**83.53**
	VCD	50.8	14.44	79.19	97.97	97.23	91.40	90.98	89.21	88.89	81.67
	DeCo	42.4	11.62	75.99	98.51	97.54	91.96	91.34	89.12	88.65	78.21
	FLB	32.6	9.57	76.32	98.53	97.60	93.99	93.31	92.31	91.72	79.05
	SHIELD	40.4	11.29	76.65	98.43	97.52	92.69	92.10	90.03	89.56	79.31
	MFCD	51.6	14.00	80.17	98.27	97.50	91.82	91.37	87.81	87.55	81.50
	HSS-DR	**32.4**	**9.11**	76.45	**98.83**	**97.87**	**94.04**	**93.34**	**92.36**	**91.74**	78.63
Qwen-VL-Chat	Vanilla	46.6	12.29	76.97	97.50	96.63	89.47	89.02	88.46	88.05	78.97
	VCD	48.2	12.16	**77.30**	98.44	97.56	90.77	90.29	88.72	88.34	**79.73**
	DeCo	41.0	10.84	75.80	98.61	97.61	91.65	91.03	89.51	89.00	77.87
	FLB	40.0	9.83	74.82	98.50	97.51	92.85	92.17	90.93	90.34	77.87
	HSS-DR	**31.0**	**7.93**	74.82	**98.70**	**97.64**	**94.31**	**93.49**	**92.86**	**92.13**	76.94
MiniGPT-4	Vanilla	36.2	10.94	**63.14**	99.02	**97.31**	92.32	91.06	92.21	90.96	**67.99**
	VCD	33.6	10.86	60.47	98.83	96.79	92.20	90.66	91.54	90.04	63.94
	DeCo	36.0	12.43	61.51	98.59	96.67	90.69	89.34	89.95	88.65	65.03
	FLB	28.4	**5.51**	60.53	**99.23**	97.30	94.05	92.50	92.93	91.45	65.46
	HSS-DR	**26.8**	8.91	58.64	99.21	97.13	**94.72**	**92.98**	**94.13**	**92.43**	63.68
InstructBLIP	Vanilla	46.2	13.20	74.30	98.18	97.19	89.40	88.88	86.86	86.47	77.62
	VCD	55.0	14.72	**74.43**	97.99	97.05	88.36	87.92	86.22	85.88	**78.21**
	DeCo	39.4	13.80	70.19	**99.42**	**98.02**	88.82	88.06	88.36	87.62	72.47
	FLB	38.6	11.62	70.32	98.77	97.57	91.52	90.74	88.50	87.88	74.75
	HSS-DR	**34.0**	**9.02**	70.45	98.88	97.64	**92.53**	**91.67**	**91.86**	**91.03**	74.32
Qwen2.5-VL	Vanilla	37.2	9.01	68.43	98.61	97.20	94.23	93.10	92.38	91.37	71.71
	VCD	36.0	8.71	**69.28**	98.85	**97.49**	93.68	92.64	92.86	91.88	**72.55**
	FLB	33.0	7.70	66.86	**99.04**	97.47	94.41	93.15	92.82	91.66	69.85
	HSS-DR	**28.2**	**6.54**	63.99	98.89	97.16	**95.13**	**93.67**	**94.12**	**92.73**	67.65
Qwen3-VL	Vanilla	54.4	9.57	**77.36**	97.91	97.03	90.80	90.30	88.49	88.09	**79.22**
	FLB	50.4	9.32	75.02	97.96	96.95	90.85	90.23	88.90	88.38	77.11
	HSS-DR	**38.6**	**6.58**	72.02	**98.75**	**97.48**	**93.56**	**92.60**	**91.89**	**91.02**	73.65

To assess general multimodal capabilities, we use MM-Vet v1 (Yu et al., 2024b), which contains 218 questions covering 200 images. Responses are scored by GPT-4.1 through the online evaluator. Further generation and evaluation details are provided in Appendix A.3.

Implementation details. Vanilla and HSS-DR use the same prompt, image preprocessing, native conversation template, and stopping conditions within each model. Both use greedy decoding with at most 512 new tokens for CHAIR and AMBER, and 1,024 for MM-Vet. The LLaVA detector is trained on 7,564 generated object candidates from 998 images, with five folds grouped by image; 2,268 candidates from a separate set of 299 images are used for calibration. CHAIR-500 is held out from detector training and calibration. HSS-DR evaluates up to four local continuations with a 12-token lookahead and permits at most three searches per response.

5.2 Experimental Results

CHAIR results. Across all six LVLMs in Table 1, HSS-DR reduces CHAIR_S by 24.2%–37.0% and CHAIR_I by 18.6%–36.7% relative to vanilla decoding. It achieves the lowest CHAIR_S on all six models and the lowest CHAIR_I on five, with FLB obtaining a lower CHAIR_I on MiniGPT-4. On Qwen-VL, HSS-DR also achieves lower CHAIR_S and CHAIR_I than FLB at the same object Recall,

improving grounding while preserving the same level of object coverage. These gains span Q-Former-based architectures and the more recent Qwen2.5-VL and Qwen3-VL backbones. Description length remains broadly unchanged under HSS-DR (Table 10), while object Recall decreases by 2.15–5.34 percentage points.

OPOPE results. HSS-DR improves Precision over vanilla decoding in all 18 model–split comparisons and $F_{0.2}$ in 16 (Table 1). Precision gains are larger on the Popular and Adversarial splits, and HSS-DR achieves the highest Adversarial Precision among the evaluated methods on all six models. These gains indicate fewer unsupported mentions of objects that are frequent in the dataset or commonly co-occur with objects present in the image. Recall decreases alongside these precision gains, consistent with the coverage changes in CHAIR. Since $F_{0.2}$ places substantially greater weight on Precision than Recall, its improvement in most comparisons indicates that the reduction in false object mentions generally outweighs the accompanying loss in object coverage under an evaluation that prioritizes Precision.

Table 2: **AMBER Generative results.**

Model	Method	CHAIR↓	Cover↑	Hal↓	Cog↓
LLaVA-1.5	Vanilla	7.39	51.74	35.06	4.22
	VCD	7.70	51.44	35.86	3.74
	DeCo	7.86	47.15	30.48	2.94
	FLB	**4.77**	48.96	**24.70**	**2.17**
	HSS-DR	5.91	**51.85**	27.69	2.58
Qwen-VL-Chat	Vanilla	6.71	53.55	31.27	2.78
	VCD	7.04	**54.19**	33.47	3.01
	DeCo	6.03	53.01	29.78	2.92
	FLB	5.58	52.46	**26.99**	2.36
	HSS-DR	**5.55**	53.81	**26.99**	**1.95**

Table 3: **MM-Vet v1 results.**

Model	Vanilla↑	HSS-DR↑	Δ
LLaVA-1.5	29.22	29.82	+0.60
Qwen-VL-Chat	46.10	46.65	+0.55
Qwen2.5-VL	66.06	66.06	+0.00

Δ is HSS-DR minus Vanilla.

AMBER results. On AMBER Generative (Table 2), HSS-DR reduces CHAIR, Hal, and Cog relative to vanilla decoding on both LLaVA and Qwen-VL while maintaining Cover. Compared with FLB, HSS-DR retains higher Cover on both models. FLB achieves lower hallucination scores on LLaVA; on Qwen-VL, HSS-DR matches FLB's Hal of 26.99 while achieving lower CHAIR and Cog. Across CHAIR and AMBER, hallucination reduction remains consistent, while the accompanying coverage changes differ: object Recall decreases on CHAIR, whereas Cover remains stable on AMBER.

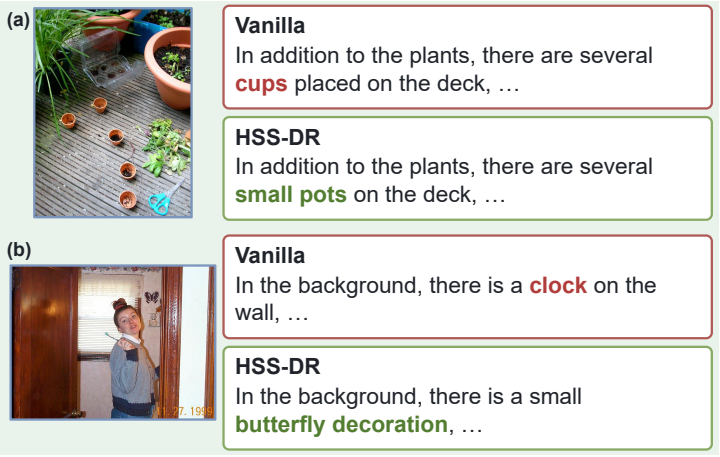


Figure 5: **Qualitative examples.** LLaVA-1.5-7B on COCO using "Please describe this image in detail." Red: hallucinated objects; green: grounded replacements; ellipses: omitted text.

MM-Vet results. Overall MM-Vet scores remain stable across LLaVA, Qwen-VL, and Qwen2.5-VL, with changes of +0.60, +0.55, and 0.00 points, respectively (Table 3). HSS-DR modifies 28, 28, and 45 of the 218 responses, respectively, indicating that its local interventions preserve performance on broader multimodal tasks.

Qualitative results. Figure 5 shows HSS-DR replacing "cups" with "small pots" and "clock" with "butterfly decoration." These examples illustrate local corrections of unsupported object mentions while leaving most of the surrounding description unchanged.

5.3 Ablation Results

Spectral features achieve the highest AUPRC among the individual feature groups (Table 4), supporting the predictive value of the spectral differences identified in Section 3. Adding spectral features to response map features and generation context increases AUPRC from 0.6743 to 0.7071, indicating that the spectral representation provides information beyond the other two inputs. The full detector achieves the highest AUROC and AUPRC, while the intermediate combinations suggest that the three feature groups provide complementary signals.

Table 4: **Detector input ablation.** Results are measured on the held-out detector split. Check marks indicate the feature groups used by each variant.

Variant	Spectral	Response maps	Context	AUROC↑	AUPRC↑
Context	×	×	✓	0.8286	0.4678
Response maps	×	✓	×	0.8872	0.5997
Spectral	✓	×	×	0.8902	0.6195
Response maps + Context	×	✓	✓	0.9165	0.6743
Spectral + Context	✓	×	✓	0.9192	0.6889
Full	✓	✓	✓	**0.9249**	**0.7071**

Table 5: **Bounded repair ablation.** Results on LLaVA-1.5 with CHAIR-500. K denotes the candidate count and H the post-mention lookahead length.

Variant	K	H	$C_S\downarrow$	$C_I\downarrow$	R↑
Vanilla	–	–	51.4	14.40	**81.47**
No lookahead	4	0	47.6	12.97	80.43
Two candidates	2	12	43.2	11.52	77.82
Full	4	12	**32.4**	**9.11**	76.45

Table 5 examines the effects of candidate expansion and lookahead. With $H = 12$, increasing K from 2 to 4 introduces two continuations initialized from alternative high-probability tokens and reduces CHAIR_S from 43.2 to 32.4 and CHAIR_I from 11.52 to 9.11. These additional candidates are selected in 270 of 463 local searches, affecting 205 images, indicating that they are frequently used alternatives to the single category-blocked greedy continuation. The expanded candidate set gives risk-guided selection more opportunities to find a continuation that satisfies both the hallucination-risk and language-model constraints. The corresponding object Recall decreases by 1.37 percentage points.

With $K = 4$, extending the lookahead horizon from $H = 0$ to $H = 12$ reduces CHAIR_S from 47.6 to 32.4 and CHAIR_I from 12.97 to 9.11. Both variants trigger search on the same 286 images, while lookahead reduces fallback to the identity continuation from 45.6% to 16.2% and increases the number of images receiving an accepted repair from 205 to 264. By evaluating object mentions introduced later in each candidate continuation, lookahead provides additional evidence for candidate selection and repair acceptance. This helps distinguish candidates that avoid the flagged category but differ in the risk of subsequent object mentions. The corresponding object Recall decreases by 3.98 percentage points.

6 Conclusion

We identify spatial spectral signatures of object hallucination in the visual hidden representations of LVLMs. Hallucinated object mentions exhibit less low-frequency energy and higher spectral centroids than grounded mentions, with the separation becoming pronounced in late decoder layers. Building on this finding, we introduce HSS-DR, which combines lightweight mention-level risk detection with bounded local repair while keeping the backbone frozen. The same detector triggers intervention at high-risk mentions and evaluates candidate continuations over a short lookahead, linking risk detection to repair selection and acceptance. Across six LVLMs, HSS-DR reduces CHAIR_S by 24.2%–37.0% relative to vanilla decoding; consistent gains on OPOPE and AMBER, together with stable MM-Vet performance, show that the method improves visual grounding while preserving general multimodal capabilities. These results demonstrate that the spatial structure of hidden visual representations provides an actionable signal for detecting and repairing object hallucinations during generation.

AI USE STATEMENT

We used generative AI tools to assist with translation, grammar polishing, and improving the readability of the manuscript. The research ideas, methodology, and experimental design were developed by the authors. The authors reviewed and revised the AI-assisted content and take full responsibility for the final manuscript.

ETHICS STATEMENT

This work aims to improve the reliability of large vision-language models by reducing object hallucination. Our experiments use existing models and publicly available benchmarks. Although HSS-DR improves visual grounding, model outputs may still contain inaccuracies or biases inherited from the underlying models. We encourage responsible use of the method and appropriate verification of generated content, particularly in applications where inaccurate descriptions could have harmful consequences. To the best of our knowledge, this work adheres to the ethical principles outlined in the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the spectral representation, hallucination detector, and risk-guided local repair procedure in Sections 3 and 4. Section 5.1 specifies the evaluated models, benchmarks, and generation settings. Appendix A provides architecture-specific adaptations, data construction and partitioning, and evaluation protocols. Appendix B details the detector feature groups, training configuration, candidate construction, and repair acceptance procedure. Additional spectral analyses in Appendix C document the data splits and aggregation procedures used to examine the consistency of the observed signatures. Upon acceptance, we will make our implementation publicly available to support independent reproduction of our results and further development of the method.

REFERENCES

N. Ahmed, T. Natarajan, and K.R. Rao. Discrete Cosine Transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974. doi: 10.1109/T-C.1974.223784.

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, Qianying Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating Object Hallucinations in Large Vision-Language Models with Assembly of Global and Local Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29915–29926, June 2025.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report, 2025b. URL https://arxiv.org/abs/2502.13923.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. DRESS: Instructing Large Vision-Language Models to Align and Interact with Humans via Natural Language Feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14239–14250, June 2024a.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object Hallucination Reduction via Adaptive Focal-Contrast Decoding. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7824–7846. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/chen24bi.html.

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 A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 49250–49267. Curran Associates, Inc., 2023. doi: 10.52202/075280-2142. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9a6a435e75419a836fe47ab6793623e6-Paper-Conference.pdf.

Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. TruthPrInt: Mitigating Large Vision-Language Models Object Hallucination Via Latent Truthful-Guided Pre-Intervention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7372–7382, October 2025.

Jiwoo Ha, Jongwoo Baek, and Jinhyun So. First Logit Boosting: Visual Grounding Method to Mitigate Object Hallucination in Large Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18241–18250, June 2026.

Yixiao He, Haifeng Sun, Pengfei Ren, Jingyu Wang, Huazheng Wang, Qi Qi, Zirui Zhuang, and Jing Wang. Evaluating and Mitigating Object Hallucination in Large Vision-Language Models: Can They Still See Removed Objects? 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. 6841–6858, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.349. URL https://aclanthology.org/2025.naacl-long.349/.

Zheng Yi Ho, Siyuan Liang, and Dacheng Tao. Understanding Hallucinations in Large Visual and Language Models. *ACM Computing Surveys*, 58(13):1–36, 2026.

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, June 2024.

Yiyang Huang, Liang Shi, Yitian Zhang, Yi Xu, and Yun Fu. SHIELD: Suppressing Hallucinations In LVLM Encoders via Bias and Vulnerability Defense. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 125123–125143, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/cae31fa913f140b3126036cc474fd461-Paper-Conference.pdf.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-Introspective Decoding: Alleviating Hallucinations for Large Vision-Language Models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 24272–24295, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3cc87f2bd3e3b4df8f9217326761c322-Paper-Conference.pdf.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in Middle Layers of Large Vision-Language Models: Interpreting, Detecting and Mitigating Object Hallucinations via Attention Lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014, June 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See What You Are Told: Visual Attention Sink in Large Multimodal Models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 87676–87703, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/da8a39bc39ae1c89dd6ebb1e3bcbb3f3-Paper-Conference.pdf.

Sohee Kim, Soohyun Ryu, Joonhyung Park, and Eunho Yang. Unveiling the Response of Large Vision-Language Models to Visually Absent Tokens. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21535–21557, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1092. URL https://aclanthology.org/2025.emnlp-main.1092/.

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, June 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 Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 23–29 Jul 2023a. URL https://proceedings.mlr.press/v202/li23q.html.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

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 Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35799–35819. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/li25ca.html.

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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Yangguang Lin, Quan Fang, Yufei Li, Jiachen Sun, Junyu Gao, and Jitao Sang. HulluEdit: Single-Pass Evidence-Consistent Subspace Editing for Mitigating Hallucinations in Large Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11086–11095, June 2026.

Bingqian Liu, Fu Zhang, Guoqing Chen, and Jingwei Cheng. Multi-Frequency Contrastive Decoding: Alleviating Hallucinations for Large Vision-Language Models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28568–28584, Suzhou, China, November 2025a. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1452. URL https://aclanthology.org/2025.emnlp-main.1452/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, June 2024.

Sheng Liu, Haotian Ye, and James Y Zou. Reducing Hallucinations in Large Vision-Language Models via Latent Space Steering. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 72402–72419, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b4008025c2182bfe16fcc8566ee14d64-Paper-Conference.pdf.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying More Attention to Image: A Training-Free Method for Alleviating Hallucination in LVLMs. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 125–140, Cham, 2025c. Springer Nature Switzerland. ISBN 978-3-031-73010-8.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards Interpreting Visual Information Processing in Vision-Language Models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 57172–57189, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/900fb3439e4968df79a6f2bfedec49cd-Paper-Conference.pdf.

Eunkyu Park, Minyeong Kim, and Gunhee Kim. HalLoc: Token-level Localization of Hallucinations for Vision Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29893–29903, June 2025.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object Hallucination in Image Captioning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

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 Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.775. URL https://aclanthology.org/2024.findings-acl.775/.

Barrett Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening Anchor Token: Decoding Strategy in Alleviating Hallucinations for MLLMs. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 27745–27776, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4537e2172096d966228d68c428392894-Paper-Conference.pdf.

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 Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 13712–13736, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/24079b91da7257cb78805262996152b8-Paper-Conference.pdf.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation, 2024a. URL https://arxiv.org/abs/2311.07397.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating Hallucinations in Large Vision-Language Models with Instruction Contrastive Decoding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15840–15853, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.937. URL https://aclanthology.org/2024.findings-acl.937/.

Jialin Wu, Wei Shi, Han Shen, Peigui Qi, Kunsheng Tang, Zhicong Huang, Binghao Wang, and Zhou Yang. REVIS: Sparse Latent Steering to Mitigate Object Hallucination in Large Vision-Language Models. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.11824.

Tsung-Han (Patrick) Wu, Heekyung Lee, Jiaxin Ge, Joseph Gonzalez, Trevor Darrell, and David Chan. Generate, but Verify: Reducing Hallucination in Vision-Language Models with Retrospective Resampling. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 65749–65777. Curran Associates, Inc., 2025. doi: 10.52202/085713-2206. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5eff08bd064f0cdd92182cdf6fd06b99-Paper-Conference.pdf.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and Mitigating Hallucination in Large Vision-Language Models via Modular Attribution and Intervention. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 51546–51568, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/8001c3568152d134d821cd46d4d84768-Paper-Conference.pdf.

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

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)*, pp. 13807–13816, June 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 Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57730–57754. PMLR, 21–27 Jul 2024b. URL https://proceedings.mlr.press/v235/yu24o.html.

Feiran Zhang, Yixin Wu, Zhenghua Wang, Xiaohua Wang, Changze Lv, Xuanjing Huang, and Xiaoqing Zheng. VIB-Probe: Detecting and Mitigating Hallucinations in Vision-Language Models via Variational Information Bottleneck. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23509–23521, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1078. URL https://aclanthology.org/2026.acl-long.1078/.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Barrett Tang, Sinan Fan, Yaomin Shen, Chen Shen, and Hao Tang. Hallucination Begins Where Saliency Drops. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and

A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 15062–15083, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/196c8da9209b1977408d8771c4e7ee56-Paper-Conference.pdf.

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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 56969–56998, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/fc625e831361cfcc82cb74224fdc66cb-Paper-Conference.pdf.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/50623630a2372839c078474efa6c0cb8-Paper-Conference.pdf.

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. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80873–80899. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/zou25e.html.

A Architecture Adaptation and Evaluation Details

A.1 Adapting Spatial Spectra to Different Architectures

All six backbones follow the response-map and spectral construction in Section 3.1. We extract visual representations during the initial prompt forward pass, apply the model's final language-model normalization to decoder input hidden states, and project them onto object directions constructed from its own output matrix. The architecture-specific step is to associate these responses with a two-dimensional spatial layout, as summarized in Table 6.

Table 6: **Spatial response maps across architectures.** Each adapter produces a spatial map before the same DCT-based feature extraction. Model names follow Section 5.1.

Model	Spatial adaptation
LLaVA-1.5	Preserve the 24×24 patch layout and crop it to the rectangular image-content region, excluding padding.
Qwen-VL-Chat	Arrange the 256 resampler responses on their native 16×16 positional grid.
MiniGPT-4 / InstructBLIP	Map the 32 Q-Former query responses back to the 16×16 image patch grid using query-to-patch cross-attention.
Qwen2.5-VL / Qwen3-VL	Recover the native rectangular grid after spatial merging, preserving the image-token spatial order and dynamic resolution.

For MiniGPT-4 and InstructBLIP, the learned queries have no direct spatial layout. We average attention heads in the final Q-Former cross-attention block, remove the visual class token, and normalize the remaining patch weights for each query. These weights distribute each signed query response over image patches; averaging across queries yields the spatial response map to which the DCT is applied. For Qwen3-VL, states are captured at decoder block inputs so that the visual information supplied by DeepStack is included at the appropriate layers.

The adapters retain the signed responses and DC component. Spatial frequency coordinates are normalized separately along the two grid axes before forming the six radial bands, allowing the same spectral statistics to be computed on rectangular grids. For the dynamic grids of Qwen2.5-VL and Qwen3-VL, the response-mass feature uses mean absolute response rather than its sum to account for varying grid sizes. Detector features use layers L24–L30 for LLaVA-1.5, Qwen-VL-Chat, MiniGPT-4, and InstructBLIP; L21–L27 for Qwen2.5-VL; and L27–L33 for Qwen3-VL, with zero-based layer indices. Each backbone has its own trained detector and risk mapping, while following the same detection and local repair procedure.

A.2 Data Construction and Partitioning

We collect object candidates during each backbone's own greedy generations. Object surface forms, including synonyms and plurals, are mapped to canonical categories using the CHAIR object vocabulary. A candidate is labeled as absent when its canonical category is missing from the image's CHAIR ground-truth object set, and present otherwise. Thus, detector examples are object candidates, and their number depends on the backbone's generation even when the input image set is shared.

Data partitions and model-selection folds are defined by image identity, keeping candidates from the same image together. The LLaVA training and calibration sample counts are given in Section 5.1. The fixed 500-image CHAIR evaluation set is excluded from detector training and calibration across all six backbones. It is shared by the final CHAIR and OPOPE evaluations, which score the same generated descriptions.

A.3 Evaluation and Generation Protocols

Prompts and generation. For CHAIR, LLaVA-1.5, MiniGPT-4, and InstructBLIP use "Please describe this image in detail." Qwen-VL-Chat, Qwen2.5-VL, and Qwen3-VL use "Describe this image in detail." This prompt choice follows our observation that Qwen-VL-Chat produced much shorter descriptions when given the same "Please describe this image in detail." prompt used for LLaVA-1.5. AMBER uses its official generative prompt, "Describe this image." MM-Vet uses the original benchmark questions. Each prompt is wrapped in the corresponding model's native conversation format. Within each backbone, Vanilla and HSS-DR share image preprocessing, prompts, and stopping conditions. Both use greedy decoding with one beam and a maximum of 512 new tokens for CHAIR and AMBER, or 1,024 for MM-Vet.

OPOPE query construction. We evaluate the object sets extracted from the CHAIR-500 descriptions. For each image, we select up to three ground-truth objects as positive queries and an equal number of absent objects as negative queries. Random negatives are sampled uniformly from absent categories; Popular negatives are selected by object frequency; Adversarial negatives are selected by co-occurrence with the positive objects, with frequency-based fallback. Frequency and co-occurrence statistics are computed from ground-truth object sets in the fixed evaluation image set. Query construction uses seed 42 and produces 2,368 queries per split. A query is predicted positive when its object occurs in the extracted description. The three splits share their positive queries, so their Recall is identical for a given set of descriptions. All methods are scored against the same queries.

AMBER and MM-Vet. AMBER evaluation uses the 1,004 generative queries, official annotations, and official scoring procedure, with metrics reported in their original units. MM-Vet v1 uses 218 questions covering 200 images. Each model answer is scored once by GPT-4.1 through the online evaluator; Overall scores and HSS-DR-minus-Vanilla differences are reported in score points. Model and baseline coverage for each benchmark is specified in the main result tables.

B Additional Details of Detection and Repair

This section supplements the LLaVA-1.5 reference method in Section 4. Architecture adaptation and evaluation protocols are described in Appendix A.

B.1 Detector Features and Training

Feature groups. Spectral features include the six band-energy proportions, the log ratio of energy outside versus inside the two lowest bands, and the spectral centroid. We summarize their variation across the selected layers using statistics such as the mean, standard deviation, extrema, and slope. Response-map features include log absolute response mass and normalized spatial entropy computed from the absolute responses. We also summarize agreement across surface forms of the same object category and their differences from a category prototype formed by averaging object directions. These image-based features are cached after the initial forward pass. Generation-context features include the probability of the preferred token, the combined probability of tokens completing the same object category, their relative support, and the generation position.

Training and triggering. We train the two gradient-boosted tree experts described in Section 4.1, using folds grouped by image. Sample weights account for repeated candidates from the same image–object pair and balance object categories and classes. Table 7 lists the reference configuration. Classifier scores are mapped to percentiles relative to grounded training candidates. During generation, the high-recall threshold determines whether a completed object candidate triggers repair. A one-token boundary check confirms that the matched surface form is a complete word before any rollback.

Table 7: **LLaVA-1.5 reference configuration.** Training and local search parameters used for the method described here.

Parameter	Value
Boosting iterations / learning rate	200 / 0.05
Maximum leaf nodes / minimum samples per leaf	15 / 20
L_2 regularization	2.0
Candidate budget K / post-mention lookahead H	4 / 12
Maximum mean language-model regret	1.0
Maximum searches per response	3
Maximum searches per object category per response	1

B.2 Constructing Constrained Continuations

At a trigger, we restore the prefix immediately before the object's first token. The identity branch continues with unconstrained greedy decoding. The object-suppressed branch masks completions of the flagged category and continues greedily under this constraint; its first step also suppresses other object completions. Two additional branches start from distinct high-probability tokens allowed by the constraint and then continue greedily. Their seeds differ from the first token of the object-suppressed branch. Special tokens and terminal punctuation are excluded as seeds.

The identity lookahead supplies a reference set of object categories. The two additional branches restrict subsequent object completions to this set, excluding the flagged category. We compare continuations over a common local horizon extending $H = 12$ tokens beyond the original mention's completion position, subject to end-of-sequence and the response length limit. Object boundaries are checked throughout this horizon before the corresponding mentions enter the risk summary.

B.3 Risk Comparison and Repair Acceptance

For a scored mention i in a candidate continuation, let r_i be its risk percentile and τ_{e_i} the trigger threshold for its expert. Its normalized excess risk is

$$\delta_i = \frac{\max(0, r_i - \tau_{e_i})}{\max(1 - \tau_{e_i}, 10^{-12})}. \tag{6}$$

The risk summary $R_{\mathrm{hall}}(b)$ in Equation 5 orders candidates by the number of unsafe mentions, the maximum excess risk, and the sum of excess risks, in that order. Empty excess-risk summaries are zero. Unsafe mentions include objects above the trigger threshold, newly introduced categories absent from the identity lookahead, and object completions for which a calibrated detector score is unavailable. Reappearance of the flagged category in a repair branch adds a further penalty to the unsafe count.

Language-model regret measures how much the selected tokens deviate from the model's preferences along each candidate's own prefix. For T_b locally generated tokens $v_1, \ldots, v_{T_b}$, let p_j be the unconstrained next-token distribution at step j and $v_j^* = \arg\max_v p_j(v)$. We use

$$R_{\mathrm{LM}}(b) = \frac{1}{T_b} \sum_{j=1}^{T_b} \left[\log p_j(v_j^*) - \log p_j(v_j) \right]. \tag{7}$$

Valid repair candidates must satisfy the regret bound in Table 7. The identity branch remains available as the fallback. Among eligible branches, the lexicographic rule in Equation 5 prioritizes the ordered risk summary, then regret; remaining ties use repetition, edit cost, and a fixed branch order.

The selected repair is committed only when it reduces unsafe mentions and satisfies the risk-reduction checks. Otherwise, the identity continuation is retained. Generation resumes from the selected continuation, with at most three searches per response and one search per object category.

C Additional Evidence of Spatial Spectral Signatures

We complement Section 3 with evidence across data splits, within individual images, and across model architectures. Throughout this section, ρ and μ follow Equation 3; window-level statistics average L24–L30 for each mention before aggregation.

C.1 Consistency across Data Splits

Table 8 compares the LLaVA-1.5 development and disjoint confirmation splits. Both show higher low-frequency energy for grounded mentions and higher centroids for hallucinated mentions. The confirmation gaps of 6.89 percentage points and 0.0516, respectively, reproduce the late-layer spectral pattern on a separate image set.

Table 8: **LLaVA-1.5 spectral differences across data splits.** G/H denote grounded/hallucinated mentions. Positive gaps in both columns indicate the spectral ordering observed in the main text; pp denotes percentage points. Statistics use L24–L30.

Split	Images	Mentions (G/H)	$\rho_G - \rho_H$ (pp)	$\mu_H - \mu_G$
Development	100	636 / 103	8.76	0.0658
Confirmation	299	1,915 / 353	6.89	0.0516

C.2 Within-Image Spectral Differences

Among the 100 development images, 52 contain both grounded and hallucinated mentions. For each image, we average the two classes separately and compute their spectral differences. Figure 6 shows that 44 of 52 images (84.6%) have higher low-frequency energy for grounded mentions, and 45 of 52 (86.5%) have higher centroids for hallucinated mentions. Thus, the aggregate class separation also appears among objects described within the same visual scene.

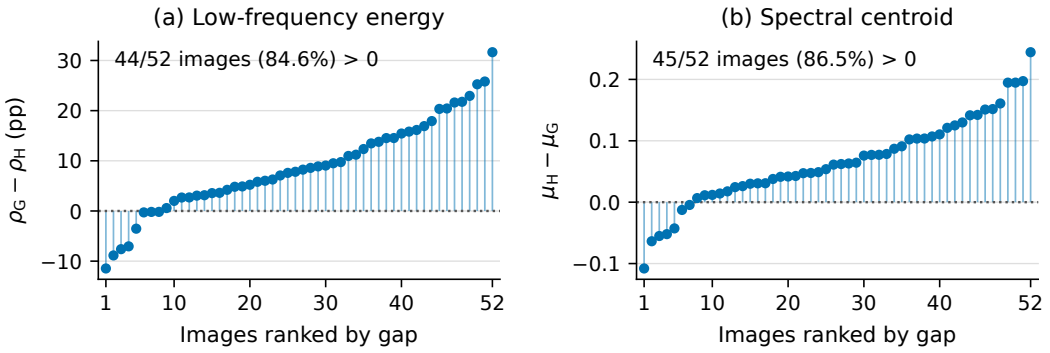


Figure 6: **Within-image spectral differences on LLaVA-1.5.** Each point represents an image containing both classes, using L24–L30; images are sorted independently in each panel. Positive gaps indicate higher grounded low-frequency energy (left) or higher hallucinated centroids (right).

C.3 Radial Cumulative Spectral Energy

For each mention and layer, we accumulate normalized DCT power up to each radial frequency on the native content grid, then average over L24–L30 and over mentions in each class. Figure 7 uses all 636 grounded and 103 hallucinated development mentions. Grounded responses accumulate energy faster at low frequencies, with the largest gap reaching 14.75 percentage points at normalized radius 0.141, further revealing their greater low-frequency concentration.

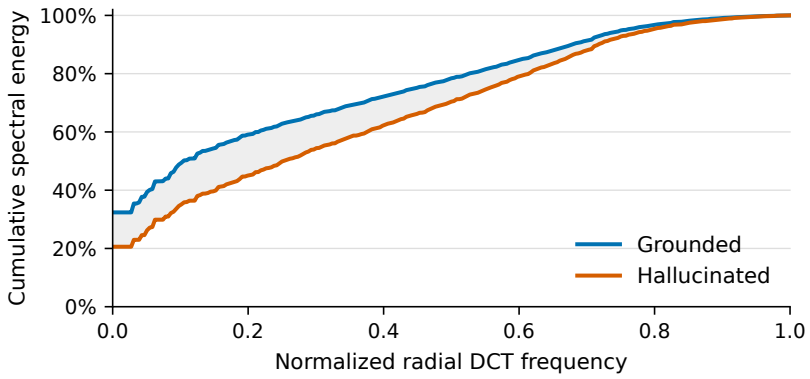


Figure 7: **Radial cumulative spectral energy on LLaVA-1.5.** Curves average normalized cumulative power over all development mentions in each class and L24–L30, using each image's native content grid. The shaded region shows the gap between the two class means.

C.4 Spectral Signatures on Qwen-VL-Chat

We apply the same spectral statistics to Qwen-VL-Chat using its 16×16 positional resampler grid, as described in Appendix A. The analysis includes all 1,584 object mentions in a 200-image set: 1,417 grounded and 167 hallucinated. Figure 8 shows the class means across its 32 decoder layers, indexed from 1 to 32. In L24–L30, grounded mentions have 7.57 percentage points more low-frequency energy, while hallucinated mentions have a 0.0564 higher centroid (Table 9). Each statistic alone reaches an AUROC of approximately 0.69 for distinguishing the two classes. The same direction of separation on Qwen-VL-Chat extends the spectral evidence to a different visual-token architecture.

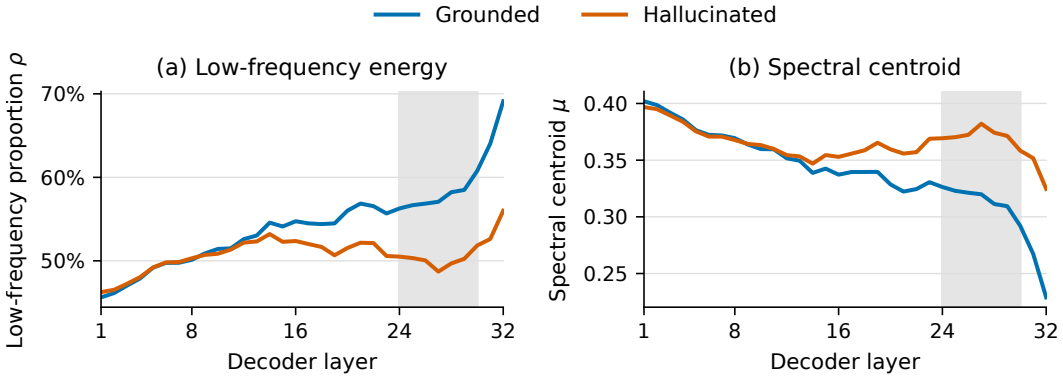


Figure 8: **Class trajectories on Qwen-VL-Chat.** Curves use all mentions in each class; shading marks L24–L30. Hallucinated mentions show lower low-frequency energy and higher centroids in this window.

Table 9: **Qwen-VL-Chat spectral statistics over L24–L30.** Means use all 1,417 grounded and 167 hallucinated mentions. AUROC treats hallucinated mentions as positive, using $-\rho$ or μ as the score.

Statistic	Grounded	Hallucinated	AUROC
Low-frequency proportion ρ	57.77%	50.20%	0.6888
Spectral centroid μ	0.3147	0.3711	0.6913

D GENERATION LENGTH AND INFERENCE TIME

We report generation length and inference time for LLaVA-1.5 on CHAIR-500, covering all seven methods in Table 1. Each statistic uses the same 500 evaluation images.

D.1 GENERATION LENGTH

Table 10 places description length alongside hallucination rates. HSS-DR maintains an average length of 101.45, nearly identical to Vanilla's 101.38, while achieving the lowest CHAIR_S and CHAIR_I among the compared methods. Its descriptions are also longer on average than those of VCD, DeCo, SHIELD, and MFCD.

Table 10: **Generation length and hallucination on LLaVA-1.5.** Avg. Length uses the CHAIR evaluator's tokenization. CHAIR scores are percentages; bold marks the lowest hallucination rates.

Method	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	Avg. Length
Vanilla	51.4	14.40	101.38
VCD	50.8	14.44	99.89
DeCo	42.4	11.62	98.80
FLB	32.6	9.57	102.60
SHIELD	40.4	11.29	99.10
MFCD	51.6	14.00	100.46
HSS-DR	**32.4**	**9.11**	101.45

D.2 INFERENCE TIME

Table 11 reports mean generation times from the evaluation runs, excluding model loading and evaluation scoring. Each method uses its corresponding decoding implementation and settings, with a maximum of 512 new tokens for all methods. Output tokens count the final model-token sequence, distinct from the CHAIR tokenization above.

Table 11: **Recorded generation time on LLaVA-1.5.** Times and output-token counts are averaged over CHAIR-500. Relative time is normalized to Vanilla.

Method	Avg. Tokens	Time / Image (s)	Relative Time
Vanilla	118.57	2.475	1.00×
VCD	117.47	4.884	1.97×
DeCo	118.11	2.918	1.18×
FLB	119.75	2.498	1.01×
SHIELD	118.36	6.447	2.60×
MFCD	118.94	7.397	2.99×
HSS-DR	118.89	3.777	1.53×

Implementation details. The original MFCD implementation batches the original, high-pass, and low-pass views together. Our implementation processes the three branches sequentially, retaining their separate KV caches and contrastive logit combination. This choice trades additional execution time for lower memory requirements; the reported MFCD time reflects this sequential implementation. SHIELD's recorded time covers its generation stage with precomputed initial descriptions.

HSS-DR takes 3.777 seconds per image, or 1.53× Vanilla's time, including spectral feature extraction, detection, and local repair. It maintains a similar output-token count while reducing hallucinations.

E Additional Qualitative Examples

Prompt: Please describe this image in detail.



LLaVA-1.5

The image features a bathroom with two toilets placed side by side, both of which are white and made of ceramic. The toilets are positioned close to each other, with one located on the left side and the other on the right side of the room.

In addition to the toilets, there is a **sink** in the bathroom, situated towards the left side of the room. The **sink** is accompanied by a **mirror**, which is placed above it. The bathroom also has a **toilet paper roll** placed on the right side of the room, near the second toilet.

LLaVA-1.5 + HSS-DR

The image features a bathroom with two toilets placed side by side, both of which are white and made of ceramic. The toilets are positioned close to each other, with one located on the left side and the other on the right side of the room.

In addition to the toilets, there is a **handrail** installed above the toilettes, providing support and accessibility for users. The bathroom appears to be clean and well-maintained, with a tiled floor and walls.



LLaVA-1.5

The image features a group of people gathered together, with a man in a suit and tie standing in the center. He is adjusting his tie, while the others watch him. There are at least nine people in the scene, with some standing closer to the man in the suit and others further away.

A few of the people in the group are wearing ties, indicating a formal or professional setting. A **handbag** can be seen placed on the ground near the center of the group. The overall atmosphere appears to be friendly and social, with the man in the suit possibly being the center of attention.

LLaVA-1.5 + HSS-DR

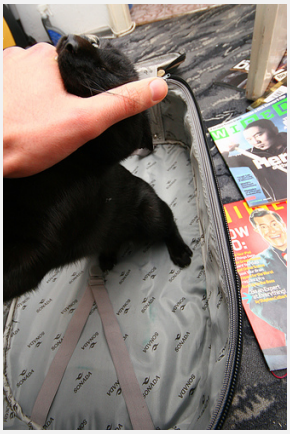
The image features a group of people gathered together, with a man in a suit and tie standing in the center. He is adjusting his tie, while the others watch him. There are at least nine people in the scene, with some standing closer to the man in the suit and others further away.

A few of the people in the group are wearing ties, indicating a formal or professional setting. A **handheld microphone** can be seen in the scene, suggesting that the gathering might be related to a public event or interview. The overall atmosphere appears to be focused and attentive, as the man in the suit adjusts his tie.

Red: hallucinated content **Green:** grounded replacement

Figure 9: **Full-response comparisons on LLaVA-1.5.** HSS-DR replaces hallucinated objects with a visible handrail (top) and a handheld microphone (bottom), while retaining the opening scene descriptions.

Prompt: Describe this image in detail.



Qwen-VL-Chat

The image features a person holding open a piece of luggage, with a small black cat sticking its head out. The person's hand is holding the edge of the suitcase, and the cat is inside, curiously exploring its surroundings. The luggage is on a **bed**, adding to the cozy atmosphere.

There are several magazines scattered around the area, including one titled "W", which might be a Women's Weekly magazine. A **remote control** is also present, indicating that the person might be getting ready for a movie night or has recently been **watching television**.

Qwen-VL-Chat + HSS-DR

The image features a person holding open a piece of luggage, with a small black cat sticking its head out. The person's hand is holding the edge of the suitcase, and the cat is inside, curiously exploring its surroundings. The luggage is on a **carpeted surface**, adding to the cozy atmosphere.

Next to the luggage, there are several magazines scattered about, indicating that the person might be packing for a trip or just arrived and is organizing their belongings. The magazines feature people on the covers, which adds variety to the scene.



Qwen-VL-Chat

The image depicts a small and dimly lit bathroom with peeling paint on the walls. The bathroom features a sink with a mirror above it and a bathtub to the left of the sink. There are several bottles placed around the bathroom, mostly on the sink area. Some of these bottles are positioned close to the sink, while others are placed near the bathtub.

A toothbrush can be seen on the right side of the sink, and a **cup** is located on the left side of the sink. The bathroom appears to be in need of updates and better lighting.

Qwen-VL-Chat + HSS-DR

The image depicts a small and dimly lit bathroom with peeling paint on the walls. The bathroom features a sink with a mirror above it and a bathtub to the left of the sink. There are several bottles placed around the bathroom, mostly on the sink area. Some of these bottles are positioned close to the sink, while others are placed near the bathtub.

A toothbrush can be seen on the right side of the sink, and **a towel is hanging over the bathtub**. The bathroom appears to be in the process of being used, with various items scattered throughout the space.

Red: hallucinated content **Green:** grounded replacement

Figure 10: **Full-response comparisons on Qwen-VL-Chat.** HSS-DR describes the carpet beneath the luggage (top) and the towel hanging over the bathtub (bottom), correcting the corresponding object mentions.

Qwen-VL-Chat prompt: Describe this image in detail.
InstructBLIP prompt: Please describe this image in detail.



Qwen-VL-Chat

The image features a group of young people riding skateboards at a cement skate park. There are at least eight boys skateboarding in the park, which also has several other people present, watching the skateboarders or simply enjoying the area. The skate park is filled with various skateboarders, creating a lively atmosphere.

In addition to the skateboarders, there are **four bicycles** in the scene. One person is **riding a bike**, and three others have their **bicycles parked** or are in the park. The park also has a car parked at the edge of the area.

Qwen-VL-Chat + HSS-DR

The image features a group of young people riding skateboards at a cement skate park. There are at least eight boys skateboarding in the park, which also has several other people present, watching the skateboarders or simply enjoying the area. The skate park is filled with various skateboarders, creating a lively atmosphere.

In addition to the skateboarders, there are **four other people** in the scene, some of them standing close to the skate park while others are a bit farther away. The park is **surrounded by trees**, which add to the pleasant outdoor setting.



InstructBLIP

The image features a person sitting on a chair, holding a half-eaten doughnut in their left hand. The doughnut appears to be a blueberry-filled pastry, with a green filling visible on the right side. The person is also holding a second doughnut in their right hand, which is not as fully consumed. The scene takes place in a park, with a bench nearby and a tree visible in the background. The person is surrounded by various objects, including a bag, **a cup**, and **a bottle**. The doughnuts are the main focus of the image, as they are being held and consumed by the person.

InstructBLIP + HSS-DR

The image features a person sitting on a chair, holding a half-eaten doughnut in their left hand. The doughnut appears to be a blueberry-filled pastry, with a green filling visible on the right side. The person is also holding a second doughnut in their right hand, which is not as fully consumed. The scene takes place in a park, with a tree visible in the background. The park setting adds to the relaxed atmosphere of the image.

Red: hallucinated content **Green:** grounded replacement

Figure 11: **Full-response comparisons on Qwen-VL-Chat and InstructBLIP.** HSS-DR replaces the bicycle description with people and trees (top), and removes additional object mentions from the doughnut description (bottom).