TOWARDS INTERPRETABLE HALLUCINATION ANALYSIS AND MITIGATION IN LVLMS VIA CONTRASTIVE NEURON STEERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision–language models (LVLMs) have achieved strong performance in visual reasoning and understanding, but can still produce hallucinations, generating content that is not adequately supported by the visual input. Many existing methods focus on the language side by intervening in model outputs or decoding processes, while the role of visual representations remains less explored. We therefore investigate hallucinations in the model's internal visual representation space. We first use sparse autoencoders (SAEs) to transform dense and entangled visual representations into a sparse, interpretable feature space. We identify two types of SAE features: always-on neurons that recur across images and image-specific neurons that respond to the content of the current image. Intervention experiments reveal that hallucinations are associated with the spurious activation of neurons unrelated to the input image. Further experiments show that enhancing image-relevant neurons while suppressing irrelevant ones can mitigate hallucinations. Based on these findings, we propose Contrastive Neuron Steering (CNS), which uses contrastive feature steering to correct spurious visual activations. CNS further applies Always-on Neuron Compensation (ANC) to offset changes in activation mass and maintain stable token-level representations. Experiments on hallucination and general-purpose benchmarks show that CNS reduces hallucinations and improves response faithfulness to the input image without compromising general multimodal performance.

1 INTRODUCTION

LVLMs have rapidly evolved into general-purpose multimodal systems with strong capabilities in visual understanding, reasoning, and interaction (Liu et al., 2023b; Dai et al., 2023; Bai et al., 2023; Zhu et al., 2023). Yet their responses can still contain visual hallucinations, producing content that is not supported by the input image. As LVLMs are increasingly used for real-world multimodal tasks, ensuring that their outputs remain grounded in visual evidence has become an important challenge (Liu et al., 2024a; Gunjal et al., 2024; Lee et al., 2018).

Prior work addresses hallucinations through visual instruction tuning, contrastive decoding, auxiliary models, and interventions on attention or hidden representations (Liu et al., 2024b; Yu et al., 2024; Leng et al., 2024; Wang et al., 2024; Favero et al., 2024; Chen et al., 2024; Yin et al., 2023; Huang et al., 2023; Zou et al., 2025). Although effective, these approaches mainly modify model outputs or decoding processes, while the internal mechanisms underlying hallucinations remain less understood. Existing analyses often emphasize language-side biases, such as the anchor pattern and text inertia (Huang et al., 2023; Liu et al., 2024d), which suggest that linguistic priors can dominate visual evidence. In contrast, systematic study of hallucinations from the perspective of LVLM visual representations remains limited. This gap is important because visual features are dense, entangled, and often polysemantic, making it difficult to determine how image evidence is organized and how its failure contributes to hallucinated outputs.

To study this gap, we focus on the internal visual representations of LVLMs. We use sparse autoencoders (SAEs) (Makhzani & Frey, 2013; Templeton et al., 2024) to decompose the dense and entangled visual features into sparse, interpretable neurons. These neurons capture visual concepts

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.

across image- and patch-level visualizations (Figures 1, 3, 4 and 12 to 14). Steering their activations can strengthen or weaken specific concepts, thereby changing the model's outputs (Figures 2, 16 and 17). Together, these properties provide a structured basis for analyzing and diagnosing hallucinations, as well as developing feature-level interventions to mitigate them.

Building on this decomposition, we analyze the organization and behavior of visual features and test their functional roles through targeted interventions. We find two complementary activation patterns: always-on neurons recur across images, whereas image-specific neurons respond selectively to the visual content of the current image. Hallucinations are associated with the spurious activation of neurons unrelated to the input image, and image-specific neurons have a stronger causal influence on hallucination-related predictions. Further interventions show that enhancing image-relevant neurons while suppressing irrelevant ones can mitigate hallucinations and correct hallucinated predictions (Figures 6, 8, 10 and 11). Together, these findings provide the basis for our feature-level approach to hallucination mitigation.

Based on these findings, we propose Contrastive Neuron Steering (CNS), an interpretable intervention applied to the visual representation before it is passed through the multimodal connector during visual prefilling. CNS uses contrastive comparison between clean and perturbed views to identify image-conditioned activation changes and correct image-specific features accordingly. Because this correction can change the total activation mass at each visual token, CNS introduces Always-on Neuron Compensation (ANC), which uses the relatively high-activation always-on neurons as a compensation buffer. Redistributing the change across these neurons limits its effect on the visual representation and stabilizes token-level activations. Together, these components provide a targeted feature-level approach to hallucination mitigation.

Our contributions are summarized as follows:

- We analyze LVLM hallucinations in a sparse, interpretable visual feature space. We identify recurring and image-specific activation patterns, link hallucinations to spurious irrelevant features, and show that targeted feature steering can mitigate them.
- We propose Contrastive Neuron Steering (CNS), which corrects image-specific activations and uses high-activation always-on neurons to compensate for changes in activation mass.
- Experiments on hallucination-focused and general-purpose benchmarks show that CNS reduces hallucinations while preserving general multimodal performance.

2 Related Work

Hallucination Mitigation in LVLMs. Despite their multimodal capabilities, LVLMs remain prone to visual hallucinations (Liu et al., 2024a; Gunjal et al., 2024; Yue et al., 2024). Prior work explores visual supervision and alignment (Liu et al., 2023a; Sun et al., 2023; Zhou et al., 2024; Liu et al., 2024b; Chen et al., 2023), contrastive or image-aware decoding (Leng et al., 2024; Wang et al., 2024; Liu et al., 2024d; Favero et al., 2024; Chen et al., 2024; Deng et al., 2024; Zhu et al., 2024; Yue et al., 2024; Li et al., 2023a), auxiliary verification and revision (Yu et al., 2024; Yin et al., 2023; Lee et al., 2023), and attention or hidden-state intervention (Huang et al., 2023; Lyu et al., 2026; 2025; Wan et al., 2025; Zou et al., 2025; Yin et al., 2025; Kang et al., 2025). These methods address hallucinations at different stages and achieve promising results. However, their interventions primarily operate at the output, decoding, or language-representation level. Existing analyses attribute hallucinations to language-side biases, including anchor patterns and text inertia (Huang et al., 2023; Liu et al., 2024d). The role of internal visual representations remains less understood. We therefore study hallucinations in this space and perform feature-level intervention before multimodal fusion.

Sparse Autoencoders. SAEs decompose dense activations into sparse latent features, providing an interpretable basis for studying and steering internal representations (Makhzani & Frey, 2013; Bricken et al., 2023; Templeton et al., 2024; Bussmann et al., 2024a;b; Nabeshima, 2024). Applications across language, vision, and multimodal models have revealed features corresponding to objects, parts, textures, and higher-level concepts (Pach et al., 2025; Thasarathan et al., 2025; Zhang et al., 2024; Lou et al., 2025; Kravets et al., 2026; Shu et al., 2025). Prior work mainly uses SAEs for feature interpretation or concept-level control. We use SAEs to transform dense visual representations into a sparse, interpretable feature space, where we analyze and diagnose hallucinations and perform feature-level interventions to mitigate them.

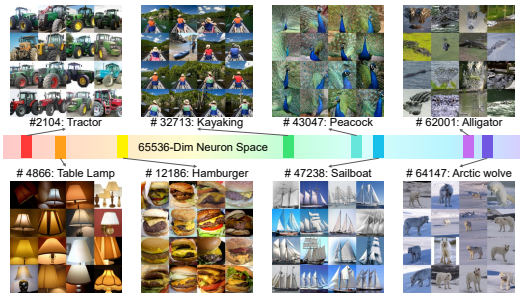


Figure 1: SAE feature patterns.

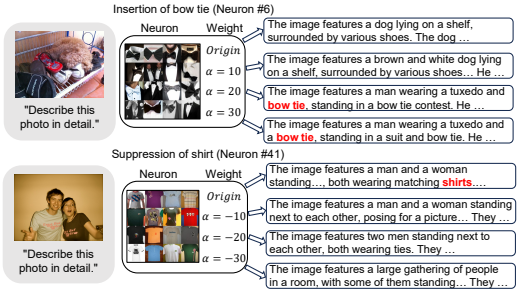


Figure 2: Feature steering changes outputs.

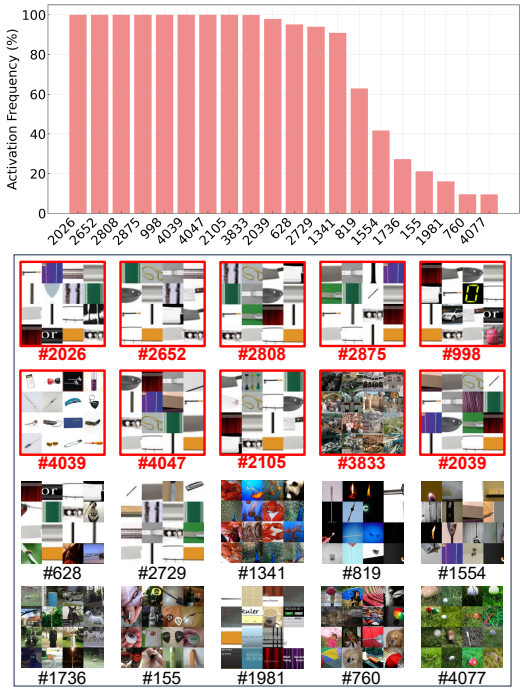


Figure 3: Image-level activation.

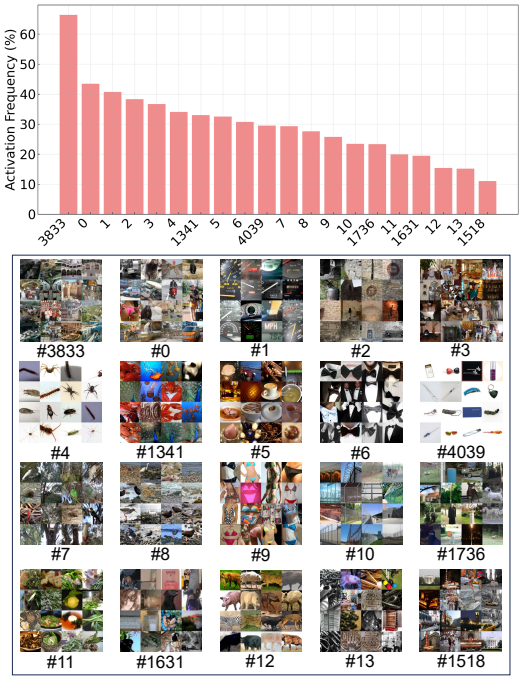


Figure 4: Patch-level activation.

3 Interpretable Hallucination Analysis and Mitigation

To understand how hallucinations arise in visual representations and how to mitigate them, we decompose dense visual features into sparse, interpretable neurons using SAEs. We analyze their organization, sensitivity to visual perturbations, and causal effects on model predictions. The results show that hallucinations can arise when visual features unrelated to the input receive abnormally high activations, motivating the design of CNS.

We train a Matryoshka BatchTopK SAE (Bussmann et al., 2024a;b; Pach et al., 2025) on visual features extracted from the frozen visual encoder of LLaVA-1.5-7B using ImageNet, and use this SAE for the LLaVA-1.5-7B analysis below.

3.1 An Interpretable and Controllable Neuron Space

Sparse decomposition for interpretability. To enable fine-grained analysis of how visual features contribute to hallucinations, we decompose dense visual representations into sparse, interpretable neurons. For an image I, the frozen visual encoder produces P visual tokens,

$$V = f_v(I) = [v_1, \ldots, v_P]^\top \in \mathbb{R}^{P \times d}. \tag{1}$$

The SAE encoder maps them to a sparse code:

$$H = g_{\text{enc}}(V) = \text{Sparse}_{K_s}\big(\text{ReLU}(VW_e^\top + b_e)\big) \in \mathbb{R}_+^{P \times m}, \tag{2}$$

where m is the number of latent features and K_s controls the sparsity. The decoder reconstructs the visual tokens as

$$\widehat{V} = g_{\text{dec}}(H) = HW_d^\top + b_d. \tag{3}$$

Each column of H is one latent feature, and $H_{p,i}$ is its activation at visual token p. We refer to these SAE latent features as neurons below. As shown in Figure 1, we can inspect the images and patches that activate each neuron most strongly to understand what visual concepts it responds to. We can also modify neuron activations and observe how predictions change (Figure 2), enabling causal analysis of their effects.

Categorizing neurons by activation patterns. To identify distinct neuron types, we measure how consistently each neuron activates across images. For an image I, let $H(I) \in \mathbb{R}_+^{P\times m}$ denote its SAE code. We summarize neuron i by its strongest token response,

$$a_i(I) = \max_{p\in\{1,\dots,P\}} H_{p,i}(I), \tag{4}$$

and denote the indices of the K_a largest responses by $\mathcal{T}_{K_a}(I)$. We compute its activation frequency on a calibration set $\mathcal{D}_{\text{cal}}$ as

$$r_i = \frac{1}{|\mathcal{D}_{\text{cal}}|} \sum_{I\in\mathcal{D}_{\text{cal}}} \mathbb{I}[i \in \mathcal{T}_{K_a}(I)]\,. \tag{5}$$

Neurons with $r_i \geq \rho$ form the always-on neuron group $\mathcal{A} = \{i : r_i \geq \rho\}$, and the remaining neurons form the image-specific neuron group $\mathcal{I} = \{1,\dots,m\} \setminus \mathcal{A}$. As shown in Figures 3 and 4, always-on neurons consistently activate across images and often respond to recurring textures or scene structure, whereas image-specific neurons activate selectively and localize to specific objects, attributes, or object parts. The appendix provides additional visualizations (Figures 12 to 14).

3.2 Diagnosing Hallucination via Feature Disruption

Measuring neuron changes under perturbation. To diagnose how hallucinations arise from internal representations, we examine how different neuron groups respond to image perturbation. Following VCD (Leng et al., 2024), we construct a perturbed view $I^n = q_t(I^c, \xi)$ for each clean image I^c, where t controls the noise level. We process both images with the same visual encoder and SAE. For a feature group $G \in \{\mathcal{A}, \mathcal{I}\}$, we measure the change in its active features by

$$R_G(I^c, I^n) = 1 - \frac{|\mathcal{T}_{K_a}^{(G)}(I^c) \cap \mathcal{T}_{K_a}^{(G)}(I^n)|}{K_a}, \tag{6}$$

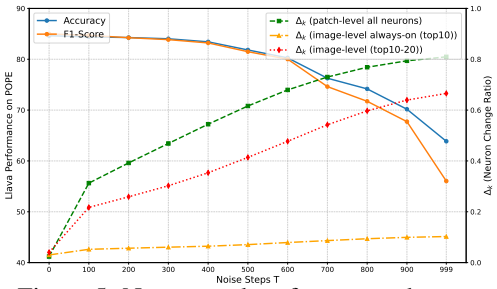


Figure 5: Neuron and performance changes.

where $\mathcal{T}_{K_a}^{(G)}$ contains the Top-K_a responses within G. We also record the signed change in each feature's image-level response:

$$\delta_i(I^c, I^n) = a_i(I^c) - a_i(I^n). \tag{7}$$

R_G measures the fraction of the active set that changes, while δ_i captures the direction and magnitude of change for individual neurons.

Visual perturbations disrupt image-specific features. We examine the effect of visual perturbations from both quantitative and qualitative perspectives. Quantitatively, we evaluate LLaVA-1.5-7B on the random split of POPE (constructed from COCO) at different noise levels. As shown in Figure 5, the image-specific neuron set changes substantially as perturbation increases, while the always-on set remains stable. POPE performance decreases over the same range, indicating that image-specific features are more sensitive to the loss of current-image evidence. At the feature level, the signed changes further show that perturbation can weaken evidence-bearing features while increasing the activation of unrelated features. Qualitatively, the camera example shows camera-related activation weakening as perturbation removes visual detail, while the generated description becomes less specific (Figure 15). Together, these results connect perturbation-induced changes in image-specific features to the emergence of hallucinations.

Quantitative Validation of Feature Effects. To quantify the causal effect of each neuron group on hallucination, we set one group to zero while leaving all other features unchanged:

$$H_{p,i}^{\setminus G} = H_{p,i}\,\mathbb{I}[i \notin G], \qquad V^{\setminus G} = g_{\text{dec}}(H^{\setminus G}). \tag{8}$$

Table 1: Performance after zeroing neurons.

Feature group	Accuracy (%) ↑	F1-score (%) ↑
No zeroing	84.63	84.99
Always-on	84.68	85.08
Image-specific	63.08	57.36
Random	84.31	84.65

We compare the always-on and image-specific groups with a random group of the same size. As shown in Table 1, removing image-specific neurons degrades hallucination benchmark performance significantly more than removing always-on or size-matched random features. The qualitative cases in Figures 6, 8, 10 and 11 demonstrate the same effect on individual concepts: suppressing an unsupported concept or enhancing a feature supported by the image changes the predicted object or attribute. Image-specific neurons encode information about the current image, whereas always-on neurons provide responses that recur across images.

3.3 Mitigating Hallucination via Feature Intervention

Feature steering corrects hallucinations. We use feature steering to examine how individual visual features affect model predictions (Figures 16 to 19). The qualitative results (Figures 6, 8, 10 and 11) reveal two related failure modes. In Figures 6 and 8, features supporting the visible content remain active, yet an unrelated feature becomes strongly activated and introduces an absent concept into the prediction. Suppressing the irrelevant feature or enhancing image-relevant features restores the correct answer. In Figures 10 and 11, object-related features are activated even though the corresponding objects are absent, leading to hallucinated object predictions. Together, these results show that targeted steering of sparse visual features can correct hallucinations by reducing irrelevant activations or strengthening image-supported ones.



Figure 6: Object hallucination under occlusion. Sheep-related features (red boxes) remain active, while dog-related feature 2480 (green boxes) is spuriously activated, producing a false "yes". Suppressing it restores "no".

Overall, these analyses show that hallucinations arise when irrelevant visual features are spuriously activated, while image-specific features encode image-dependent evidence and are more sensitive to perturbations. Always-on neurons tend to have larger activations and rank near the top across inputs, providing a stable reference for ANC. These findings motivate CNS, which corrects image-specific activations while stabilizing the visual representation.

4 Method

Our analysis in Section 3 suggests that hallucinations can arise from erroneous activations in the visual representation. Image-relevant responses may be weakened, while unsupported concepts may be spuriously activated. Feature-level interventions show that enhancing relevant responses and suppressing erroneous ones can improve visual grounding. Since always-on neurons have limited causal effects, we treat them as a stable background buffer rather than direct correction targets.

These observations motivate Contrastive Neuron Steering (CNS). As shown in Figure 7, CNS compares a clean image with a matched perturbed view to obtain an image-conditioned correction signal. It applies this signed difference to image-specific features and then uses Always-on Neuron Compensation (ANC) to offset the resulting change in token-level activation mass. The corrected representation is decoded using the original inference path.

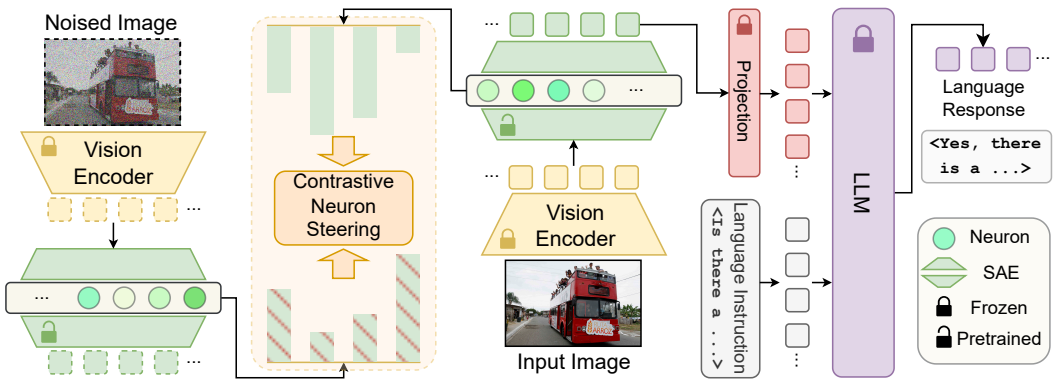


Figure 7: Overview of CNS.

4.1 Contrastive Neuron Steering

Let $\mathcal{A}$ and $\mathcal{I}$ denote the always-on and image-specific neuron groups defined in Section 3.1; the partition is fixed during inference.

Contrastive feature comparison. To obtain an image-conditioned correction direction, CNS constructs a matched perturbed view for each clean image I^c: $I^n = q_t(I^c, \xi)$, where q_t is a diffusion-based perturbation operator. Both views are processed by the same visual encoder and SAE:

$$H^u = g_{\text{enc}}(f_v(I^u)), \qquad u \in \{c, n\}, \qquad H^u \in \mathbb{R}_+^{P \times m}. \tag{9}$$

Their signed contrast is

$$\Delta H = H^c - H^n. \tag{10}$$

A positive entry corresponds to a response weakened by the perturbation, and a negative entry to one strengthened by it.

Correction of image-specific neurons. To focus the intervention on image-dependent content while leaving the always-on group unchanged, CNS applies the signed contrast to image-specific neurons starting from the clean code:

$$\overline{H}_{p,i} = \text{ReLU}\big(H^c_{p,i} + \lambda\, \mathbb{I}[i \in \mathcal{I}] \Delta H_{p,i}\big), \qquad \lambda \geq 0. \tag{11}$$

Positive differences restore weakened image evidence, while negative differences attenuate perturbation-amplified responses. Other coordinates remain at their clean values.

Always-on neuron compensation (ANC). Although the update is restricted to image-specific neurons, it can change the activation mass of each visual token. To maintain a stable representation scale, we define this change as

$$\Delta m_p = \sum_{i \in \mathcal{I}} \overline{H}_{p,i} - \sum_{i \in \mathcal{I}} H^c_{p,i}. \tag{12}$$

ANC compensates for this change by uniformly redistributing it across always-on neurons:

$$\widetilde{H}_{p,i} = \begin{cases} \overline{H}_{p,i} & \text{if } i \in \mathcal{I}, \\ \overline{H}_{p,i} - \frac{\Delta m_p}{|\mathcal{A}|} & \text{if } i \in \mathcal{A}, \end{cases} \tag{13}$$

where $|\mathcal{A}|$ is the size of the always-on neuron set.

Reconstruction and native inference. After this compensation, the corrected code is decoded and passed through the original multimodal connector:

$$\widetilde{V} = g_{\text{dec}}(\widetilde{H}) = \widetilde{H} W_d^\top + b_d, \qquad \widetilde{X} = f_{\text{con}}(\widetilde{V}, X). \tag{14}$$

The language model generates from $\widetilde{X}$ with the same inference settings as the baseline; the perturbed view is used only to compute the correction.

Table 2: POPE results (Accuracy and F1 score), averaged over three splits.

	LLaVA-1.5-7B		InstructBLIP		Qwen-VL		LLaVA-1.5-13B		LLaVA-NeXT-7B	
Method	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑
Vanilla	80.61	81.86	77.83	79.36	84.07	82.02	79.65	81.60	81.98	80.15
VCD	80.63	82.05	78.63	79.68	83.91	82.05	79.96	81.80	81.83	80.05
M3ID	81.75	82.91	79.21	80.53	84.83	82.93	-	-	-	-
OPERA	84.21	83.55	79.56	81.36	-	-	-	-	-	-
PAI	81.02	82.54	78.49	79.83	84.43	82.44	80.19	82.12	82.07	80.16
VAF	81.66	83.05	79.32	80.68	85.19	83.25	80.86	82.62	82.88	81.02
VAR	82.24	84.08	79.99	81.28	85.80	83.80	81.69	83.28	83.55	81.85
PADE	83.33	84.94	80.84	82.24	86.66	84.96	82.84	84.35	84.29	83.26
CNS (Ours)	84.12	85.24	81.28	82.65	87.18	85.34	83.21	84.89	85.18	83.64

Table 3: CHAIR results (C_S/C_I: sentence-/instance-level; maximum length 128).

	LLaVA-1.5-7B		InstructBLIP		Qwen-VL		LLaVA-1.5-13B		LLaVA-NeXT-7B	
Method	C_S ↓	C_I ↓	C_S ↓	C_I ↓	C_S ↓	C_I ↓	C_S ↓	C_I ↓	C_S ↓	C_I ↓
Vanilla	55.1	16.4	57.4	17.6	52.1	16.7	50.4	14.7	30.2	10.9
VCD	54.4	16.6	60.7	18.0	50.4	17.2	49.6	14.3	29.8	10.6
M3ID	56.6	15.8	62.3	18.2	49.8	17.4	-	-	-	-
Woodpecker	57.6	16.7	60.8	17.6	51.8	16.3	-	-	-	-
HALC	51.0	14.8	53.8	15.7	49.6	15.4	-	-	-	-
OPERA	51.6	14.2	54.2	14.8	48.6	14.6	-	-	-	-
PAI	53.1	15.1	54.2	15.6	49.1	15.6	49.4	14.1	29.4	10.4
VAF	50.1	14.2	53.4	15.1	48.7	14.4	48.6	13.4	28.4	9.6
VAR	49.6	14.1	52.3	14.7	48.4	13.9	48.2	13.2	28.2	9.2
PADE	48.6	13.7	51.8	14.2	47.8	13.4	47.3	12.6	27.8	8.9
CNS (Ours)	47.2	12.8	50.1	13.4	46.4	12.6	46.2	11.9	26.8	8.4

5 EXPERIMENTS

Benchmarks. We evaluate CNS on two categories of benchmarks. The hallucination-focused benchmarks include POPE (Li et al., 2023b), CHAIR (Rohrbach et al., 2018), Hallusion-Bench (Guan et al., 2024), and AMBER (Wang et al., 2023), covering binary object hallucination, open-ended object hallucination, and fine-grained visual faithfulness. The general-purpose benchmarks include MME (Fu et al., 2023), MM-Vet (Yu et al., 2023), VizWiz (Gurari et al., 2018), and LLaVA-Bench-in-the-wild(LLaVA-Wild) (Liu et al., 2023b), covering visual perception, reasoning, and real-world multimodal understanding. Following prior work (Leng et al., 2024; Wan et al., 2025; Lyu et al., 2026), we use the conventional metrics and standard splits for the remaining benchmarks.

Evaluated LVLMs. The main comparison considers six representative open-source LVLMs, including the classic LLaVA-1.5-7B (Liu et al., 2023b), InstructBLIP (Dai et al., 2023), Qwen-VL (Bai et al., 2023), the larger LLaVA-1.5-13B, the newer LLaVA-NeXT-7B (Liu et al., 2024c), and the more recent Qwen3.5-4B model (Qwen Team, 2026). This selection covers diverse visual encoders, connector designs, language backbones, and model scales, providing a comprehensive evaluation setting. Unless otherwise specified, we use LLaVA-1.5-7B for all experiments.

Baselines. We compare CNS with Vanilla and representative methods from three intervention interfaces: contrastive decoding (VCD (Leng et al., 2024), PAI (Liu et al., 2024d), and M3ID (Favero et al., 2024)), auxiliary-model methods (HALC (Chen et al., 2024) and Woodpecker (Yin et al., 2023)), and internal-signal intervention (PADE (Lyu et al., 2026), OPERA (Huang et al., 2023), VAF (Yin et al., 2025), and VAR (Kang et al., 2025)).

Implementation Details. We apply CNS to six representative LVLMs with different architectures and scales. We decompose the visual representations produced by their frozen visual encoders using a common SAE setup across models. Specifically, we use a standard Matryoshka BatchTopK SAE (Bussmann et al., 2024a;b; Pach et al., 2025) with Matryoshka groups

Table 4: HallusionBench.

Method	fACC ↑	qACC ↑	Easy ↑	Hard ↑
Vanilla	17.9	8.13	36.0	36.7
VCD	13.9	11.4	33.0	34.7
OPERA	16.2	5.49	37.6	35.4
PADE	18.1	11.56	37.9	37.4
CNS	19.4	11.23	38.6	39.2

Table 5: AMBER.

Method	CHAIR ↓	Cov. ↑	Hall. ↓	Cog. ↓
Vanilla	7.8	51.0	36.4	4.2
VCD	7.5	50.8	36.2	4.1
OPERA	7.3	49.6	32.0	3.5
PADE	7.1	51.8	31.4	3.4
CNS	6.8	53.1	29.8	3.5

Table 6: Qwen3.5-4B.

Method	POPE		CHAIR		MME
	Acc. ↑	F1 ↑	C_S ↓	C_I ↓	Total ↑
Vanilla	88.30	87.19	23.80	8.28	2125.96
VCD	88.52	87.43	23.15	8.06	2131.24
CNS	89.15	88.05	21.82	7.35	2142.18

Table 7: General multimodal benchmarks (LLaVA-1.5-7B).

Method	VizWiz ↑	MME ↑	LLaVA-Wild ↑	MM-Vet ↑
Vanilla	50.00	1864.68	64.80	31.1
VCD	44.90	1872.87	63.21	30.2
OPERA	50.76	1784.34	64.31	32.0
PADE	52.08	1892.12	65.92	32.4
CNS (Ours)	52.84	1908.45	67.28	33.6

$\{0.0625\omega, 0.1875\omega, 0.4375\omega, \omega\}$, expansion factor 64, and maximum sparsity $K_s = 20$. The SAE is trained with Adam (Kingma & Ba, 2017) for 10^5 steps using batch size 4096 and learning rate $16/(125\sqrt{\omega})$. Following prior work (Leng et al., 2024), we use sampling-based decoding. All experiments are conducted on NVIDIA A6000 GPUs.

5.1 Main Experimental Results

Results on Hallucination Benchmarks. As shown in Tables 2 to 6, CNS improves over Vanilla across the evaluated hallucination benchmarks and achieves the best results on most reported metrics, while remaining competitive with representative baselines across model architectures and scales. The improvements span binary object verification, open-ended object hallucination, and fine-grained visual consistency. Together, these evaluations show that CNS improves both discrete object-level decisions and free-form responses across different hallucination settings and model families. These gains partly reflect CNS's ability to suppress concepts that are spuriously activated during visual encoding, before they propagate to multimodal decoding.

Results on General Multimodal Benchmarks. We next assess whether hallucination mitigation comes at the cost of broader multimodal abilities. As shown in Table 7, CNS maintains strong performance across a range of general multimodal benchmarks covering visual perception, reasoning, and real-world multimodal understanding. These results show that the intervention improves hallucination robustness without sacrificing the model's broader capabilities. Together with the results above, they suggest that CNS improves the model's use of image evidence while preserving its overall multimodal performance across these tasks.

5.2 Ablation Studies

Effect of Proposed Components. We perform a component ablation study of CNS on LLaVA-1.5-7B using POPE and CHAIR.

We compare four variants: Vanilla; dense contrastive steering (Dense CS); sparse contrastive steering without ANC (Sparse CS); and the full CNS. Dense CS applies the same clean–perturbed contrastive difference before the sparse decomposition, whereas Sparse CS applies it to the SAE code. The comparison between Sparse CS and CNS isolates the contribution of ANC: the former uses the sparse contrastive correction alone, while the latter additionally compensates for the resulting change in token-level activation mass.

Table 8: Component ablation.

Variant	POPE F1 ↑	$CHAIR_S$ ↓	$CHAIR_I$ ↓
Vanilla	81.86	55.1	16.4
Dense CS	82.14	54.6	16.1
Sparse CS (w/o ANC)	83.78	50.3	14.2
CNS (Ours)	85.24	47.2	12.8

Steering Strength λ. We study the sensitivity to the steering strength by varying λ while holding the perturbation timestep fixed.

As shown in Table 9, performance first improves as the steering strength increases and then declines when the intervention becomes too strong. Weak steering provides an insufficient correction, whereas excessive steering can over-correct the visual representation. A moderate strength therefore offers the best trade-off between visual correction and preserving the model's original behavior. The sweep therefore identifies a stable operating region for applying steering without excessively altering the original representation.

Table 9: Steering strength sensitivity.

Setting	POPE F1 ↑	CHAIR_S ↓	CHAIR_I ↓
$\lambda = 0.5$	83.85	49.6	14.1
$\lambda = 1.0$	85.24	47.2	12.8
$\lambda = 1.5$	84.91	47.8	13.2
$\lambda = 2.0$	83.64	49.1	14.6

Perturbation Timestep t. We further vary the perturbation timestep while holding the steering strength fixed. As shown in Table 10, the results follow a similar trend: small timesteps produce a weak contrastive signal, while large timesteps introduce overly strong perturbations that may obscure reliable visual evidence. An intermediate timestep provides the most informative contrast and achieves the best overall balance between correction and representation fidelity. This behavior is consistent with using the timestep as a contrastive probe that reveals unreliable responses while retaining useful visual information.

Table 10: Perturbation timestep sensitivity.

Setting	POPE F1 ↑	CHAIR_S ↓	CHAIR_I ↓
$t = 100$	83.12	50.4	14.9
$t = 300$	84.36	48.7	13.6
$t = 500$	85.24	47.2	12.8
$t = 700$	84.68	48.1	13.4
$t = 900$	83.89	49.3	14.2

5.3 Inference Efficiency

We evaluate inference cost on a single NVIDIA A6000 GPU using the same generation settings for all methods. As shown in Table 11, CNS is more efficient than methods that require repeated full decoding or additional large expert models. CNS needs only one additional visual-encoder pass and the lightweight SAE encoder and decoder. In contrast, methods that require multiple complete decoding passes or auxiliary models incur substantially greater computational cost. CNS therefore provides a more efficient inference path while retaining strong hallucination mitigation.

Table 11: Inference efficiency.

Method	Latency ↓	GPU Memory ↓
Vanilla	3.19 s (×1.00)	14945 MB (×1.00)
VCD	6.44 s (×2.02)	15749 MB (×1.05)
M3ID	6.52 s (×2.04)	15575 MB (×1.04)
OPERA	22.64 s (×7.10)	22706 MB (×1.52)
Woodpecker	9.76 s (×3.06)	22199 MB (×1.49)
HALC	20.64 s (×6.47)	23084 MB (×1.54)
CNS (Ours)	3.47 s (×1.09)	15226 MB (×1.02)

6 Conclusion

We investigate visual hallucinations in LVLMs through their internal visual representations. Sparse, interpretable features reveal always-on neurons that recur across images and image-specific neurons that respond selectively to the current visual content. We find that hallucinations are often associated with the spurious activation of irrelevant visual features: suppressing them or enhancing relevant ones can correct hallucinated predictions. Based on these findings, we propose Contrastive Neuron Steering (CNS), which compares clean and perturbed views to identify image-conditioned changes in the visual representation and correct image-specific features before multimodal fusion. Always-on Neuron Compensation (ANC) uses relatively high-activation always-on neurons to offset changes in token-level activation mass. Across hallucination and general-purpose benchmarks, CNS reduces hallucinations while preserving broader multimodal performance.

Limitations and Future Work. CNS currently uses fixed perturbation and steering settings, which may not be optimal for every input. Future work could extend the visual-space intervention to a more general multimodal setting, investigate cross-layer feature interactions, and analyze hallucinations from complementary perspectives.

AI USE STATEMENT

In preparing this submission, the authors used a generative AI assistant to polish the language and improve the readability of the manuscript. The authors reviewed all AI-assisted changes and take responsibility for the final content of this work, including its text, claims, and artifacts.

REFERENCES

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2, 2023.

Bart Bussmann, Patrick Leask, and Neel Nanda. Batchtopk sparse autoencoders. *arXiv preprint arXiv:2412.06410*, 2024a.

Bart Bussmann, Patrick Leask, and Neel Nanda. Learning multi-level features with matryoshka saes. *AI Alignment Forum*, 2024b.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024.

Zhiyang Chen, Yousong Zhu, Yufei Zhan, Zhaowen Li, Chaoyang Zhao, Jinqiao Wang, and Ming Tang. Mitigating hallucination in visual language models with visual supervision. *arXiv preprint arXiv:2311.16479*, 2023.

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. *NIPS*, 36:49250–49267, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, 2024.

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

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *CVPR*, pp. 14303–14312, 2024.

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

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 *CVPR*, pp. 14375–14385, 2024.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *AAAI*, 2024.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pp. 3608–3617, 2018.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*, 2023.

Nick Jiang, Amil Dravid, Alexei Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. In *NeurIPS*, 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. *arXiv:2503.03321*, 2025.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Alexey Kravets, Da Li, Chuan Li, Da Chen, and Vinay P. Namboodiri. Interpretability transfer from language to vision via sparse autoencoders. In *International Conference on Machine Learning*, 2026.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. Hallucinations in neural machine translation. *OpenReview*, 2018.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*, 2023.

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 *CVPR*, pp. 13872–13882, 2024.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *ACL*, pp. 12286–12312, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023b.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023a.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023b.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pp. 26296–26306, 2024b.

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

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *arXiv preprint arXiv:2407.21771*, 2024d.

Hantao Lou, Changye Li, Jiaming Ji, and Yaodong Yang. Sae-v: Interpreting multimodal models for enhanced alignment. In *ICML*, 2025.

Guangtao Lyu, Xinyi Cheng, Chenghao Xu, Qi Liu, Muli Yang, Fen Fang, Huilin Chen, Jiexi Yan, Xu Yang, and Cheng Deng. Revealing perception and generation dynamics in lvlms: Mitigating hallucinations via validated dominance correction. *arXiv:2512.18813*, 2025.

Guangtao Lyu, Qi Liu, Chenghao Xu, Jiexi Yan, Muli Yang, Xueting Li, Fen Fang, and Cheng Deng. Revealing and enhancing core visual regions: Harnessing internal attention dynamics for hallucination mitigation in lvlms. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15214–15230. Association for Computational Linguistics, 2026.

Alireza Makhzani and Brendan Frey. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Noa Nabeshima. Matryoshka sparse autoencoders. *AI Alignment Forum*, 2024.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models. *arXiv preprint arXiv:2504.02821*, 2025.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *arXiv preprint arXiv:2503.05613*, 2025.

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

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos Derpanis. Universal sparse autoencoders: Interpretable cross-model concept alignment. *arXiv preprint arXiv:2502.03714*, 2025.

Zifu Wan, Ce Zhang, Silong Yong, Martin Q Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. *arXiv preprint arXiv:2507.00898*, 2025.

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

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

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

Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *CVPR*, pp. 14625–14634, 2025.

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. *arXiv preprint arXiv:2310.16045*, 2023.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *CVPR*, pp. 12944–12953, 2024.

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. *arXiv:2308.02490*, 2023.

Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *arXiv preprint arXiv:2402.14545*, 2024.

Kaichen Zhang, Yifei Shen, Bo Li, and Ziwei Liu. Large multi-modal models can interpret features in large multi-modal models. *arXiv preprint arXiv:2411.14982*, 2024.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024.

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

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*, 2024.

Xin Zou, Yizhou Wang, Yibo Yan, Sirui Huang, Kening Zheng, Junkai Chen, Chang Tang, and Xuming Hu. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. In *ICML*, 2025.

A Detailed Feature-Level Hallucination Cases

The following case studies make the feature-level mechanisms discussed in Section 3 directly observable. We organize them around two related failure modes. In the first, image-supported evidence remains present, but an irrelevant visual feature becomes strongly activated and introduces an absent attribute or object. In the second, a feature for an absent object is activated without reliable support from the image. For each group, we inspect the associated sparse visual features and then steer them to determine how they affect the prediction. These examples complement the quantitative benchmark and feature-zeroing results in the main paper. More importantly, they separate two questions that are difficult to answer from the final response alone: whether the correct visual evidence is still represented, and whether an unsupported concept has been introduced by another active feature.

Image-supported evidence with spurious activations. The examples in Figures 8 and 9 show that hallucination does not always require the disappearance of the correct visual evidence. In the black-apple example, apple-related features 3085 and 1941 remain active, but the red-related feature 2836 receives a strong response. Another unsupported feature, 2061, is associated with a red crane and is also activated. The resulting prediction is "red" even though the image contains a black apple. Suppressing either 2836 or 2061 restores "black," while amplifying the apple-specific features produces the same correction. These interventions indicate that the error is linked to the abnormal activation of irrelevant visual factors rather than to a complete loss of apple evidence.

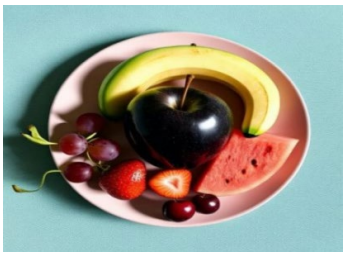
The sheep example exhibits the same pattern at the object level. Features associated with the visible sheep remain active, but partial occlusion is accompanied by a strong response from dog-related feature 2480. This unsupported response leads to a false "yes" for the dog-existence question. Once feature 2480 is suppressed, the answer changes to "no." Together, the two examples show that an irrelevant feature can distort the output while the features supporting the actual image content remain available for recovery.

Unsupported object activations. The examples in Figures 10 and 11 isolate a more direct form of object hallucination. No tennis ball is visible in the tennis scene, yet the tennis-ball feature 3541 is strongly activated and the model describes a ball near the racket. Suppressing 3541 removes both the hallucinated mention and the positive answer to the ball-existence question. A similar effect appears in the dining scene: the image contains food and drinks but no fork, while fork-related feature 7010 is activated and causes the model to answer "yes." Suppressing this feature restores the negative answer. These examples show that a single sparse feature can introduce an object-level concept even when the corresponding object is absent from the image.

Perturbation sensitivity of spurious activations. The perturbation experiments provide a broader view of the same mechanism. Under stronger perturbations, dog- and fork-related features move upward in the activation ranking, while the red-related response becomes stronger in the black-apple example. This behavior is consistent with the aggregate analysis in Figure 5, where image-specific features change more substantially under perturbation than the always-on group. The qualitative cases therefore connect individual feature changes to the broader observation that perturbations can weaken image-dependent evidence and promote unsupported visual concepts.

Implications for CNS. Across these examples, hallucinations arise either from weakened support for the image content or from the activation of irrelevant or absent concepts. Feature-level interventions make these alternatives distinguishable: relevant features can be enhanced when their evidence is weak, while unsupported features can be suppressed when they introduce an incorrect concept. CNS extends this local analysis by deriving the correction from clean–perturbed feature differences across the active image-specific features, with ANC compensating for the resulting change in activation mass.

The common interpretation is that hallucination is not simply a failure in which the visual encoder loses all information about the image. In several examples, evidence for the correct object or attribute remains active, but the prediction is redirected by an unrelated response. This view explains why both suppressing an unsupported feature and strengthening a relevant feature can repair the same type of output error. It also motivates a correction that operates on the internal visual representation before multimodal decoding, rather than relying only on post-hoc changes to the generated text.



Q: What is the color of the apple?
A: The color of the apple is red.
Suppress the 2836 neuron associated with strawberries.
Q: What is the color of the apple?
A: The color of the apple is black.
Enhance the 3085, 1941 neurons associated with apple.
Q: What is the color of the apple?
A: The color of the apple is black.



Figure 8: **Attribute hallucination from spurious feature activation.** The image contains a black apple, but the model answers "red" when asked for its color. A red-related feature (ID 2836; green boxes) becomes abnormally strong, while apple-related features remain active (red boxes). Feature 2061, associated with a red crane, is also spuriously activated. Suppressing either irrelevant feature 2836 or 2061, or amplifying the apple-specific features 3085 and 1941, restores the answer "black." The example shows that image-supported evidence can remain available while unrelated visual features receive stronger responses.



Q: Please describe the image in detail.
A: The image features two people walking down...
They are accompanied by a **sheep**, which ….

Q: Is there a **dog** in the image?
A: Yes, …

Suppress the 2480 neuron associated with dog.
Q: Is there a **dog** in the image?
A: No, …



Figure 9: **Object hallucination from a spurious activation under occlusion.** The sheep-related features (red boxes) support a generally correct description, but partial occlusion also activates a dog-related feature (green boxes) even though no dog is present. This produces a false "yes" to the dog-existence question. Suppressing feature 2480 removes the unsupported dog concept and changes the answer to "no."



Q: Please describe the image in detail.
A: The image captures a tennis player in action on a clay court. The player is in the middle of a swing, holding a tennis racket and preparing to hit the ball. The **tennis ball is visible** in the air, close to the player's racket.

Q: Is there a **tennis ball** in the image?
A: Yes, …

Suppress the 3541 neuron associated with tennis ball .

Q: Is there a **tennis ball** in the image?
A: No, …

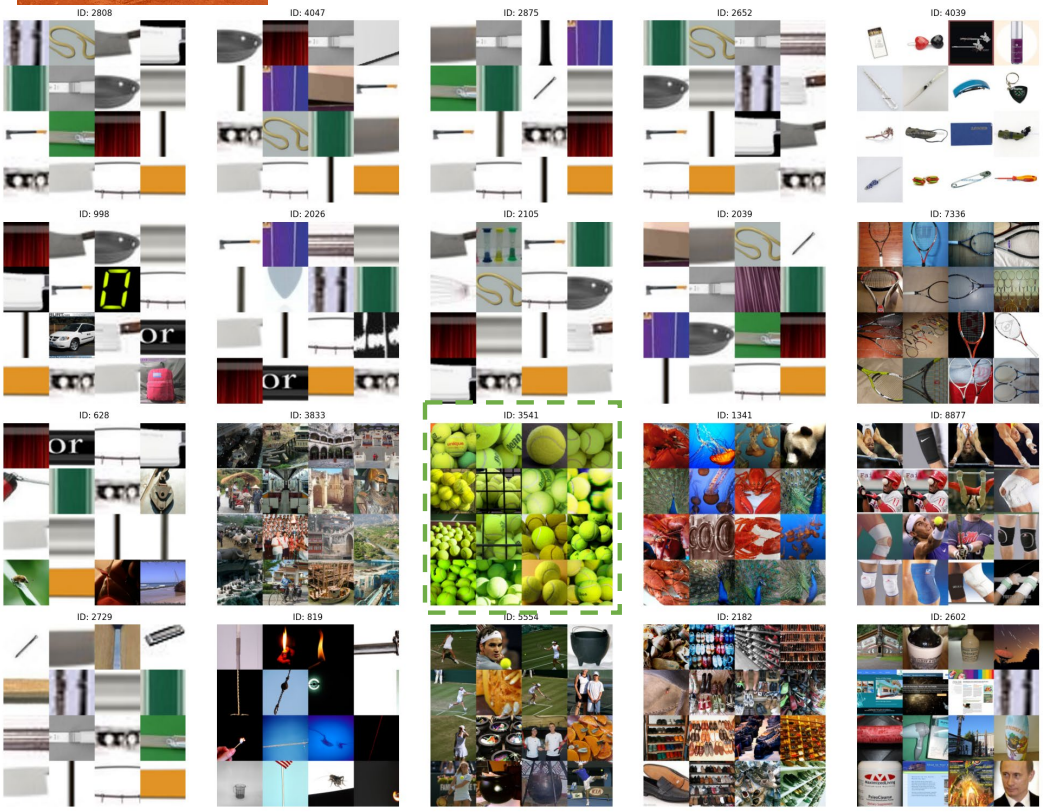


Figure 10: **Object hallucination from unsupported feature activation.** The image shows a tennis player swinging a racket on a clay court, but no tennis ball is visible. Nevertheless, the model describes a ball near the racket and answers positively to a ball-existence query. The tennis-ball feature 3541 is strongly activated despite the absence of a ball. Suppressing this feature removes both the hallucinated mention and the positive existence response.



Figure 11: **Object hallucination from contextual feature activation.** The dining scene contains food and drinks but no fork. The model nevertheless mentions a fork and answers "yes" to the corresponding question. A fork-related feature (ID 7010) is spuriously activated despite the absence of a fork. Suppressing this feature removes the unsupported concept and restores the answer "no," showing how contextual visual content can activate a semantically related but unsupported object.

B Additional Visualizations of Sparse Visual Features

The purpose of this section is to establish that the sparse coordinates used in the analysis have visually inspectable content beyond the specific hallucination examples. This provides a qualitative check before we interpret feature activations or draw conclusions from their interventions. We first ask whether the strongest activations of a feature reveal a coherent visual pattern, and whether this pattern remains meaningful across different images. This check is useful because a large activation alone does not explain what a feature represents or whether it is supported by the current image.

Figure 12 presents additional SAE features, ranging from object-associated features for corn (#14174), oranges (#46469), and dogs wearing Christmas hats (#61697) to a structural feature for spiral- or fan-shaped patterns (#62747). The examples span objects, textures, and repeated visual structures, illustrating that the sparse code covers more than one semantic level. Some features respond to a recognizable object, whereas others are better described by shape, texture, or a recurring visual pattern. This diversity is important for the later hallucination analysis. An unsupported concept may be introduced by a feature that is meaningful in general, but unrelated to the specific image under consideration. Inspecting the feature across multiple examples helps separate a stable visual association from an incidental activation.

The top-activating images and patches also provide a practical way to inspect a feature before using it for intervention. The image-level examples suggest what visual pattern a coordinate responds to, while the localized patches show which part of the image supports that response. These two views are complementary. An object label may describe the broad association, but the patches often reveal whether the activation is driven by the object itself, a specific attribute, or a related texture.

These visualizations also show that a feature need not correspond to one exact concept to be useful. A feature may respond to related appearances, such as several views of the same object or a family of similar textures, while remaining more localized and interpretable than a dense visual token. Such features should therefore be understood as recurring visual associations rather than fixed human-readable labels. This level of interpretability is sufficient for comparing clean and perturbed activations, identifying visual factors that change together, and testing how a small set of coordinates affects the corresponding output concept.

For the hallucination analysis, these visualizations provide an independent way to interpret a feature before drawing conclusions from its effect on the model output. A feature can be associated with a broad visual family rather than a single object, but its top-activating patches still reveal which visual patterns are consistently represented. This makes it possible to distinguish a feature that is plausibly related to the current scene from one that is merely activated by an incidental correlation. Visual coherence is not by itself evidence of a causal role. Instead, it gives the intervention a meaningful interpretive starting point and helps connect a change in activation to a change in the predicted concept.

The broader implication is that interpretability here is not limited to assigning one human-readable label to every coordinate. A feature can encode a visual family or a recurring structure and still be useful for intervention if its activation changes in a predictable direction with the visual evidence. The sparse representation therefore serves as an interface between visual evidence and model behavior. Visual inspection suggests what a feature may represent, activation changes show how it responds to the input, and intervention tests whether that response matters for the output. This combination supports the feature-level diagnosis and mitigation developed in the following sections.

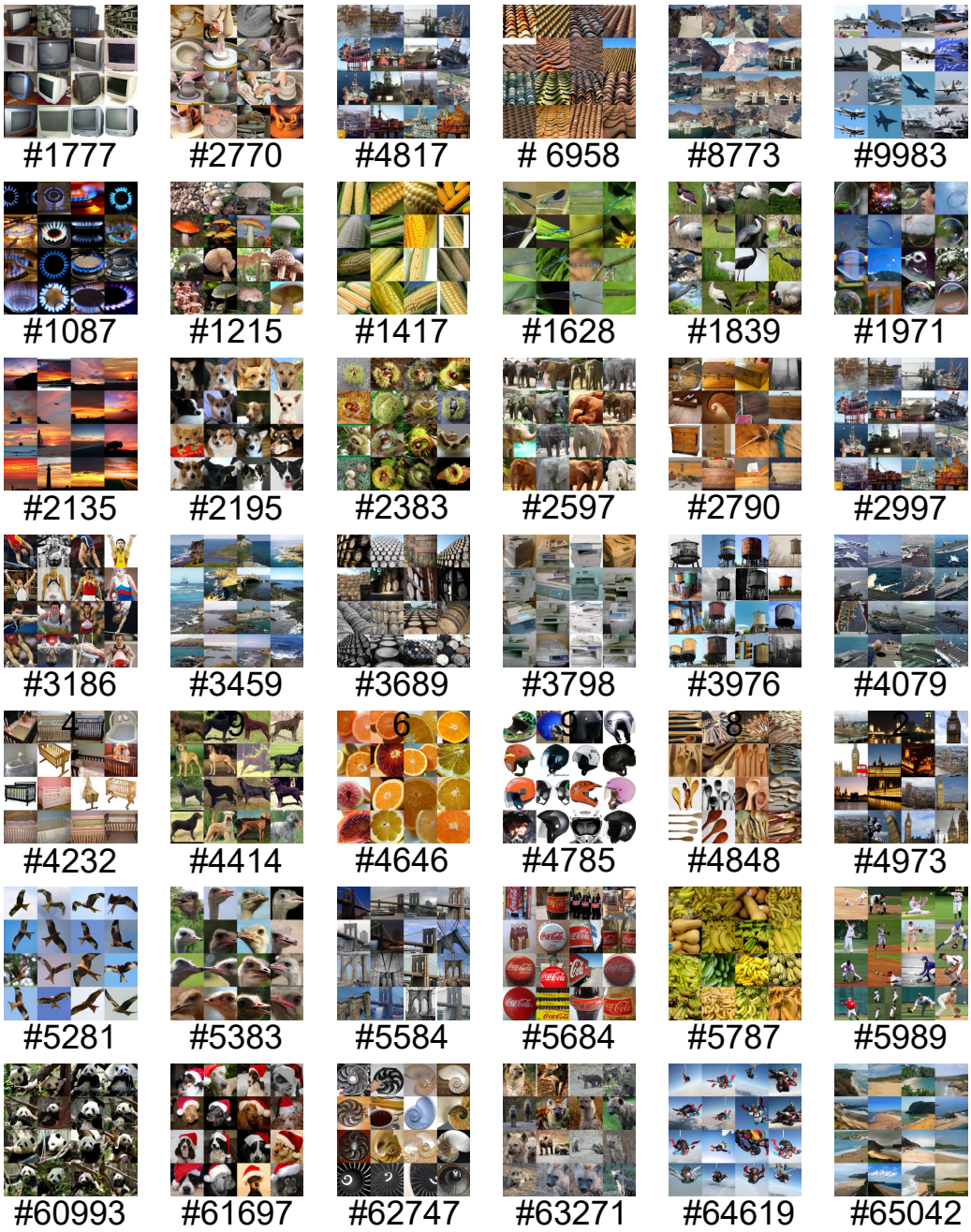


Figure 12: **Additional SAE feature visualizations.** The examples range from object-associated features, including corn #14174, oranges #46469, and dogs wearing Christmas hats #61697, to a structural feature for spirals or fan-like shapes #62747. The top-activating images and localized patches reveal both concrete and abstract visual semantics, illustrating the diversity and interpretability of the learned sparse representation. The examples also show that sparse visual factors can operate at different semantic scales, from recognizable objects to shape- and texture-level regularities. Such visualizations provide the evidence needed to interpret a feature before using its activation for analysis or steering.

C IMAGE- AND PATCH-LEVEL FEATURE ORGANIZATION

We complement the quantitative analysis in Section 3 with image-level and patch-level visualizations. These views answer two related questions: whether a feature recurs across inputs and which visual region supports its response. Image-level recurrence motivates the distinction between always-on and image-specific neurons, while patch-level localization provides evidence for the object, attribute, or region associated with an activation. A large response for one image does not make a feature persistent, and a recurring feature may still respond to a local texture. We therefore use recurrence and localization together rather than treating either as a complete semantic definition.

Image-level recurrence. Figure 13 highlights features with consistently high activation across images. The recurring features often respond to common textures, small visual patterns, or broad scene structure rather than one unique object. In the analyzed representation, their activations can reach 10–80, whereas many top-40 responses lie around 5–15. These magnitudes are descriptive; the group is defined by recurrence, as formalized in Section 3.1. Thus, a large activation alone is not sufficient to identify an always-on feature. The frequency view and the top-activating images are considered together to separate features that recur across inputs from features that respond strongly only to particular images.

The image-level view also clarifies how the always-on group is identified in practice. We examine both how often a feature appears among the strongest responses and which images produce those responses. A feature that is frequent but visually unrelated to the current scene is treated as a persistent background factor, while a feature that appears only for a subset of images is more informative about image-dependent content. This operational distinction does not imply that persistent features are meaningless. It only determines which features are used as a stable reference and which features are examined for image-conditioned changes.

For CNS, always-on features provide a stable reference for the activation-mass change compensated by ANC, whereas image-specific features provide the main coordinates for the clean–perturbed correction. Their recurrence does not imply that they are non-semantic; it makes them less suitable as direct indicators of what is unique to the current image. The partition therefore describes how a feature is used by the intervention rather than assigning a fixed semantic status to every recurring feature.

Patch-level localization. Figure 14 shows where selected features activate at the patch level. Many responses localize to distinctive textures, object parts, or small regions that carry fine-grained visual evidence. This helps explain why a feature changes across images or when a relevant region is occluded or corrupted by noise, and provides a visual check on its interpretation. It also clarifies why image-level activation can be selective even when the same feature appears in several examples: the activation may depend on a particular part, texture, or spatial configuration rather than on the whole image.

Patch-level evidence is especially useful when a feature has a broad or ambiguous image-level association. The activated patches can show whether the response is tied to an object, an attribute, a boundary, or a recurring local texture, and indicate what evidence may be weakened by a perturbation. In this way, patch localization links a change in a sparse coordinate to a concrete change in the input, which is difficult to establish from image-level activation statistics alone.

Together, these views support the partition used by CNS: recurrence identifies a stable group for activation-mass compensation, while localization helps identify image-specific features whose changes carry information about the current scene. Recurrence indicates whether a coordinate is stable across inputs, whereas localization indicates whether its response is tied to image-dependent evidence. Together they provide an interpretable partition for CNS: ANC accounts for stable activation backgrounds, while contrastive steering focuses on feature changes specific to the current image. Neither view alone proves a causal role, but their combination clarifies why the two feature groups are treated differently in the intervention.

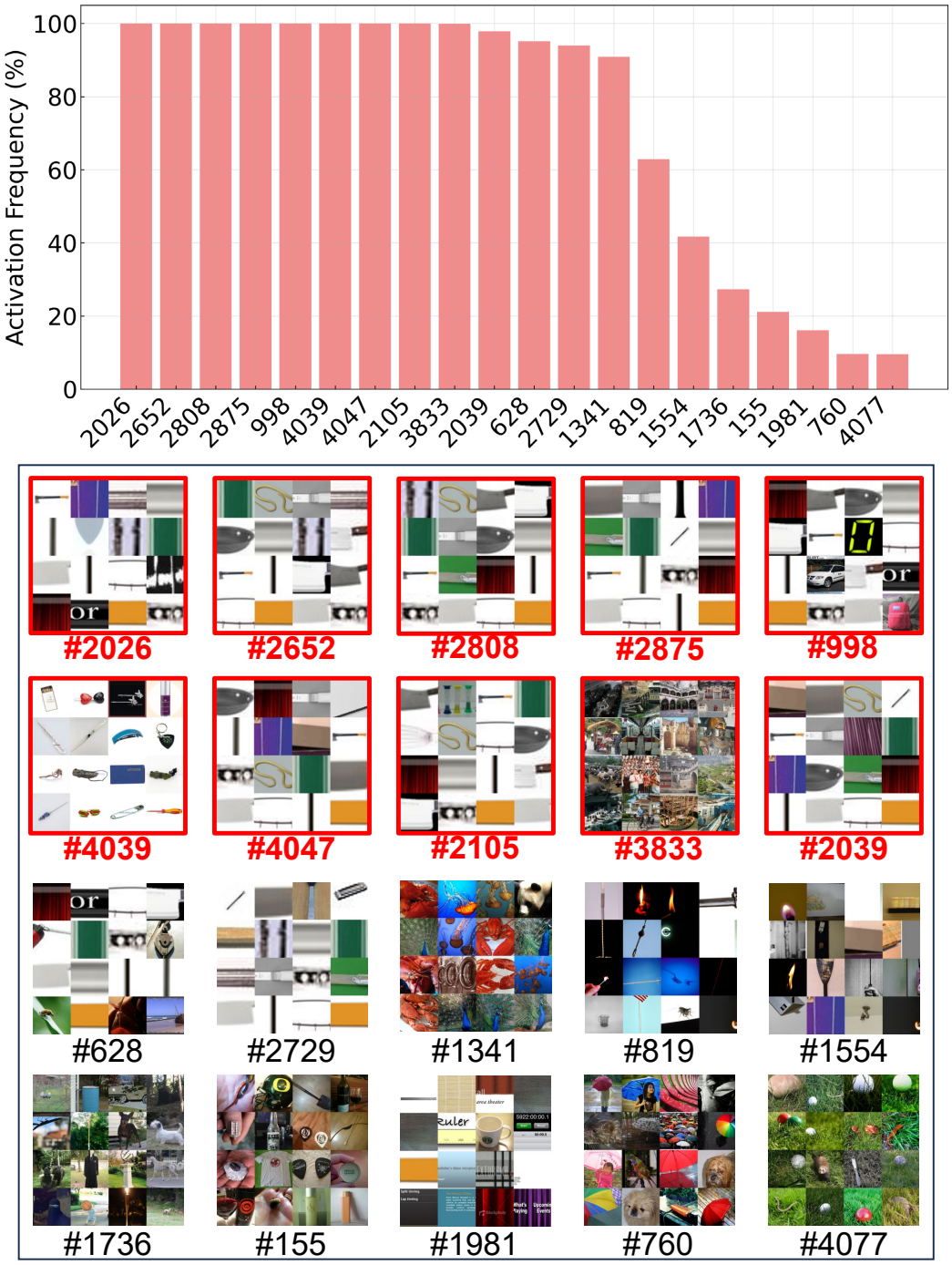


Figure 13: **Image-level SAE feature analysis.** The top panel shows activation frequency across images, and the bottom panel shows the top-activating images for each feature. Red boxes mark features that recur across images and are therefore assigned to the always-on group. These features often respond to recurring textures, small visual patterns, or broad scene structure. Their recurrence makes them useful as a stable reference for ANC, whereas the less recurrent features provide a more image-specific basis for contrastive steering. The figure also emphasizes that recurrence, rather than activation magnitude alone, defines the always-on group.

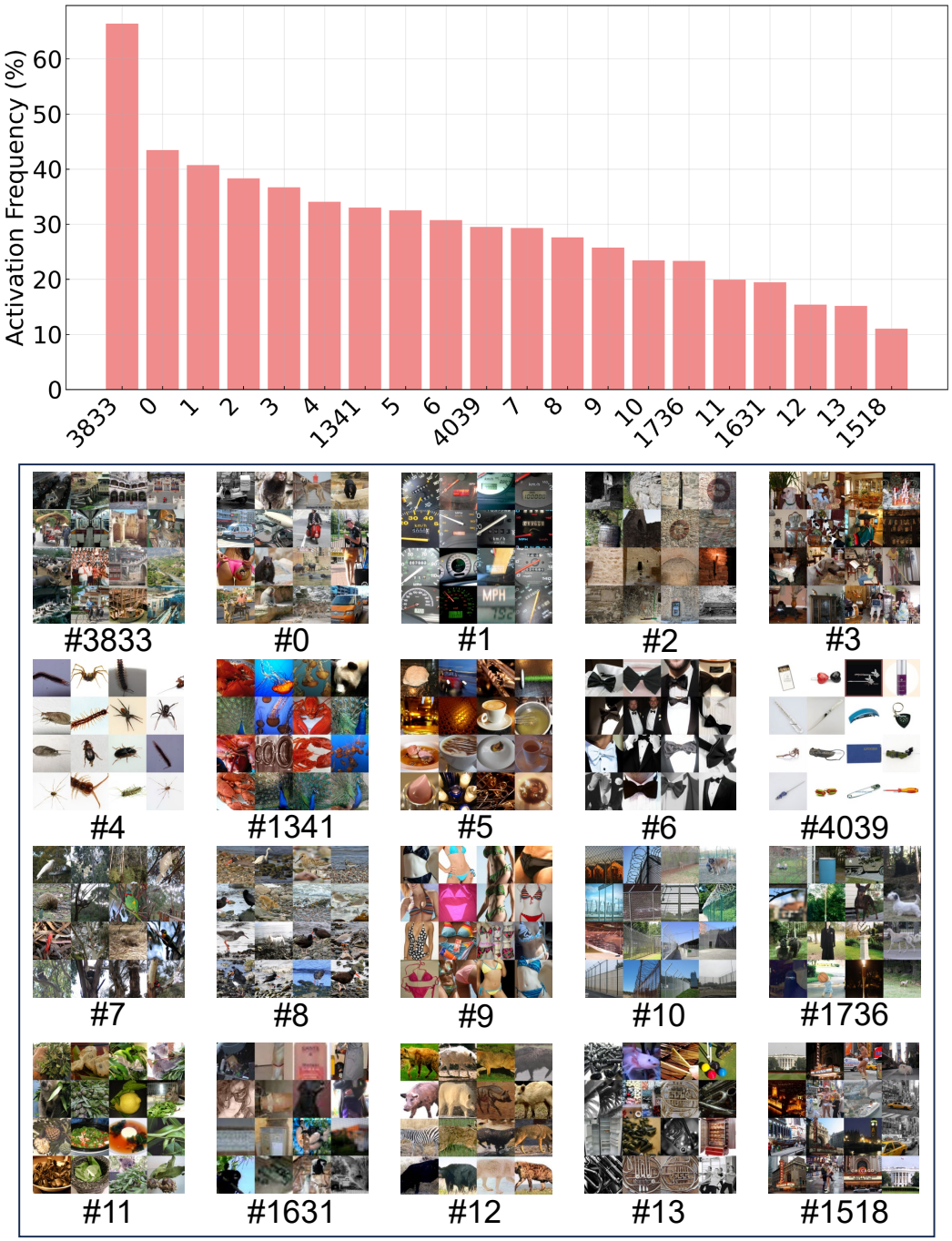


Figure 14: **Patch-level SAE feature analysis.** Localized activations reveal image-specific visual concepts, including distinctive textures, object parts, and small regions of an image. The patch-level view provides spatial evidence for interpreting a feature and helps distinguish scene-specific visual content from features that recur broadly across inputs. It further shows how a feature's response is tied to a particular region or visual pattern, which helps explain why its activation changes when that region is occluded, corrupted, or replaced by unrelated content.

D PERTURBATION EFFECTS ON IMAGE-SPECIFIC FEATURES

In Section 3, we measure how image perturbations change internal visual features and how these changes relate to hallucination benchmark performance (Figure 5). Figure 15 provides a qualitative view of the same process. The example connects perturbation strength, an interpretable feature activation, and the content of the output. Because the object remains conceptually the same while detail is progressively lost, it helps distinguish a gradual loss of visual evidence from a change caused only by final decoding. It also shows that the reliability of visual evidence can change before the model produces an explicit hallucinated description, making the feature trajectory an early signal of representational degradation.

The image contains a camera. With little or no perturbation, the model describes a "black Konica Minolta camera with a large lens." As the perturbation becomes stronger, the camera feature weakens and the output first becomes "camera with a large lens" and eventually omits the camera altogether. The sequence illustrates a gradual loss of visual specificity rather than a single abrupt decoding error: fine-grained identity is lost first, followed by the broader object concept. At the same time, perturbation can increase the relative prominence of weakly grounded alternatives, which is why the feature changes are relevant to both missed evidence and hallucinated content.

The trajectory is consistent with a hierarchy of visual information. The camera identity is lost before the broader camera concept, suggesting that perturbation first removes fine-grained evidence and only later disrupts the general object signal. In a hallucination setting, this redistribution of representational support can make an unrelated feature relatively more prominent even when its absolute activation changes only moderately.

This example also illustrates why image-specific features are useful for diagnosis. Their responses vary with the evidence available in the current image, while recurring background features can remain comparatively stable. The sparse code lets us track the camera-related response and relate its change to the generated output. The signed clean–perturbed difference used by CNS exploits this signal: it restores responses weakened by perturbation and attenuates responses that are amplified by it. ANC then compensates for the resulting change in token-level activation mass.

Together with the aggregate results, the example suggests that perturbation effects are structured rather than random. Different parts of the representation become unreliable at different stages, so the model may first produce a less specific but still plausible description and later become more susceptible to unsupported concepts. The example does not imply that every hallucination follows the same camera-like trajectory; it illustrates instead how an output change can be linked to an interpretable change in visual features. This supports viewing hallucination as a representational imbalance in which image-dependent support is weakened and unrelated responses become relatively more influential. This perspective also motivates tracking the direction of feature changes, rather than using perturbation only as a binary test of whether the final answer is correct.



Figure 15: **Perturbation-induced loss of a camera feature.** The image contains a camera, but increasing perturbation progressively weakens the camera-related activation. The output correspondingly changes from "black Konica Minolta camera with a large lens" to "camera with a large lens" and finally drops the camera description. The example links a localized change in the sparse visual representation to a gradual loss of semantic detail and motivates the clean–perturbed feature comparison used by CNS. The progression also shows that fine-grained identity can be lost before the broader object concept, illustrating why tracking feature changes is more informative than inspecting only the final decoded answer.

E FEATURE-LEVEL STEERING AND CONCEPT CONTROLLABILITY

We next examine how visual concepts are distributed across SAE features and how this affects feature-level steering. These cases are controlled probes of the intervention interface rather than additional benchmark results. We test whether a concept is carried by one dominant feature or a distributed group, how surrounding content affects steering, and whether coordinated updates are more effective than changing one feature alone.

Concept-specific sensitivity. The comparison in Figure 16 shows that different visual concepts can have different intervention sensitivities within the same scene. Suppressing chair-related features removes the chair from the output at approximately $\alpha = -30$, whereas dog-related features require around $\alpha = -100$. The larger magnitude is consistent with a more distributed dog representation: different features may encode breeds, poses, scale, or contextual cues. Suppressing one of these features can therefore leave other parts of the dog representation active. The result suggests that the strength of a steering intervention reflects the organization of a concept in the feature space, rather than only the visibility of the object in the image. The edited feature should therefore not be viewed as the only carrier of the concept; other features may preserve it when one coordinate is changed.

Context dependence. The comparison in Figure 17 shows that the same type of intervention also depends on the surrounding scene. In a relatively simple, bird-dominated image, amplifying one dog-related feature with $\alpha = 50$ is enough for the dog concept to appear in the output. In the more complex scene, the corresponding insertion requires approximately $\alpha = 500$. The larger value reflects the need for the inserted signal to remain effective in the presence of more active objects and attributes. A feature that is controllable in a simple context may therefore have a weaker behavioral effect when the representation contains richer visual content. This context dependence is important for interpreting manual steering results: the same feature and the same nominal intervention strength do not necessarily have the same effect across images. It also motivates deriving the correction from the current image and its perturbed counterpart rather than using a fixed edit for every input.

Distributed feature steering. The complex-scene results in Figures 18 and 19 further show why distributed updates can be useful. For insertion, changing one dog-related feature requires $\alpha = 500$ before a visible change appears, whereas jointly steering three related features with $\alpha = 20$ each produces the concept with a much smaller per-feature magnitude. The same pattern holds for suppression: a single feature requires approximately $\alpha = -100$, while three related features at $\alpha = -10$ each remove the concept more effectively. These results indicate that the relevant information is not uniformly localized in one coordinate. Coordinated updates can follow the structure of the learned representation and avoid placing an excessively large intervention on a single feature. Distributed updates help in both directions, for introducing a concept and for removing it.

Implications for CNS. These probes support the image-conditioned, multi-feature correction used by CNS. A fixed single-feature edit can be unreliable because its effect depends on the concept, the surrounding scene, and how the concept is distributed. CNS instead uses the clean–perturbed comparison to identify active image-specific features whose responses have changed, applies the correction across that set, and uses ANC to compensate for the aggregate activation change. The results explain why CNS operates on a structured group of features rather than assuming that every concept has one identifiable neuron.

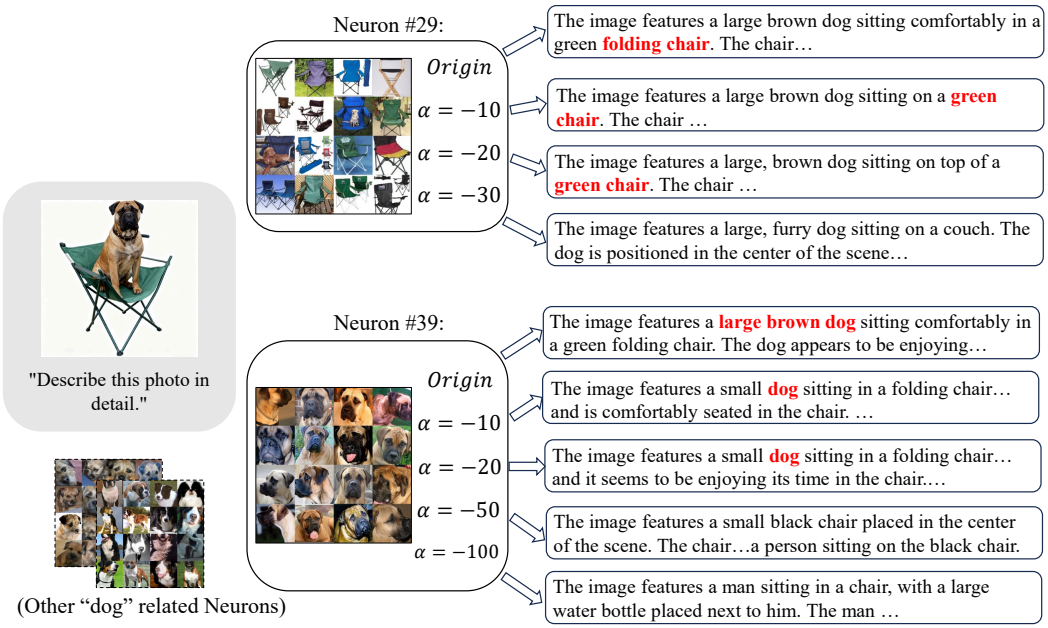


Figure 16: **Multi-concept suppression.** Suppressing chair-related features removes chairs with a moderate intervention (approximately $\alpha = -30$), while dog-related features require a larger magnitude (approximately $\alpha = -100$). The difference is consistent with a more distributed dog representation involving multiple breeds, poses, and contextual cues, and illustrates why a single-feature edit may leave a concept partially active.

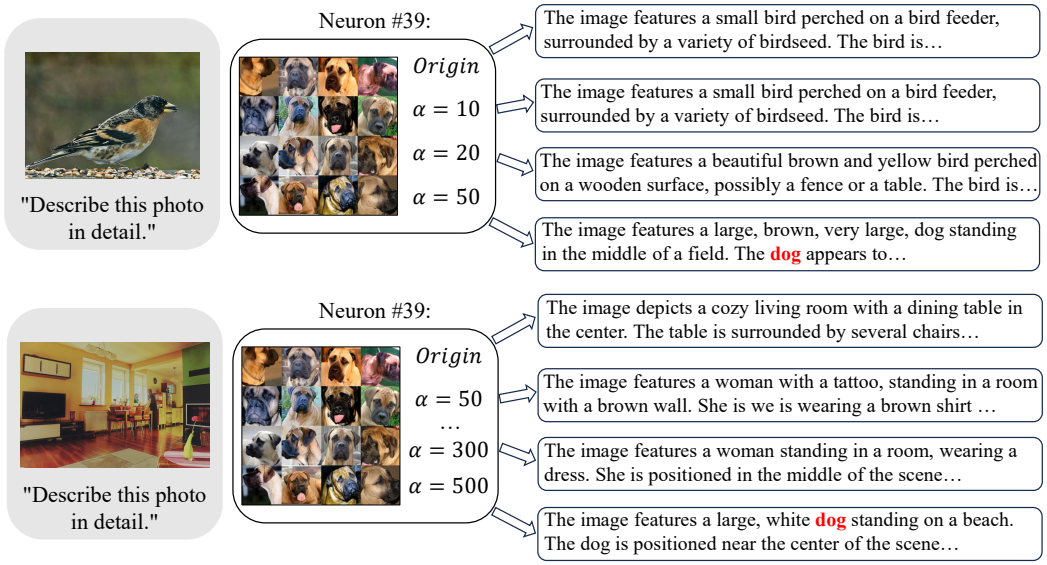


Figure 17: **Concept insertion in simple and complex contexts.** Amplifying a dog-related feature introduces the concept into a simple, bird-dominated scene at moderate strength ($\alpha = 50$), while a more complex scene requires a substantially larger intervention ($\alpha = 500$). The comparison shows that surrounding visual content affects the controllability of a concept.

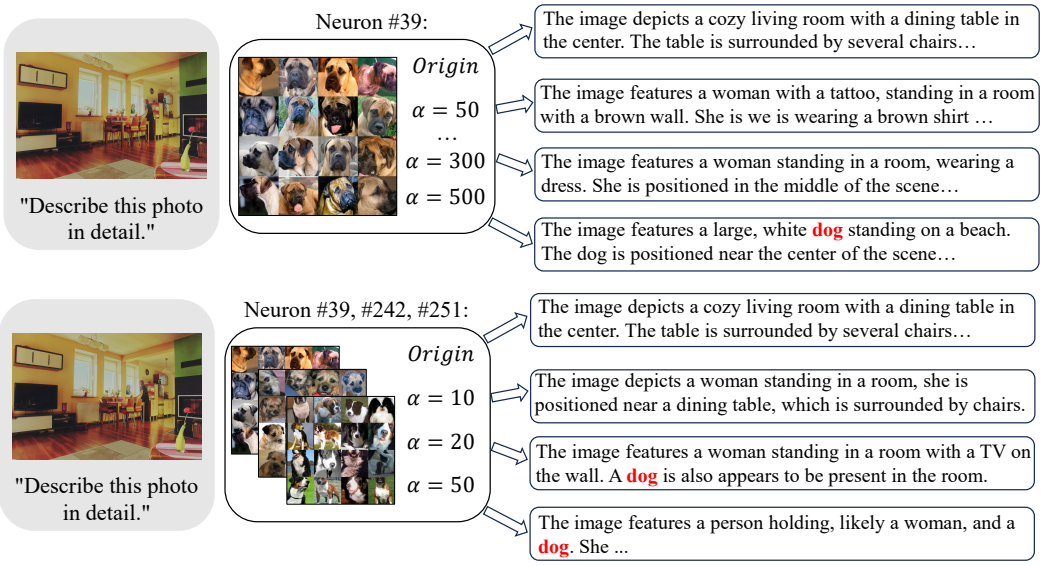


Figure 18: **Concept insertion in a complex context.** (a) Steering one dog-related feature requires a large intervention ($\alpha = 500$) to produce a visible effect. (b) Coordinated updates of three related features with $\alpha = 20$ each produce the concept with a much smaller per-feature magnitude. The result demonstrates that multi-feature steering better matches the distributed representation of the concept.

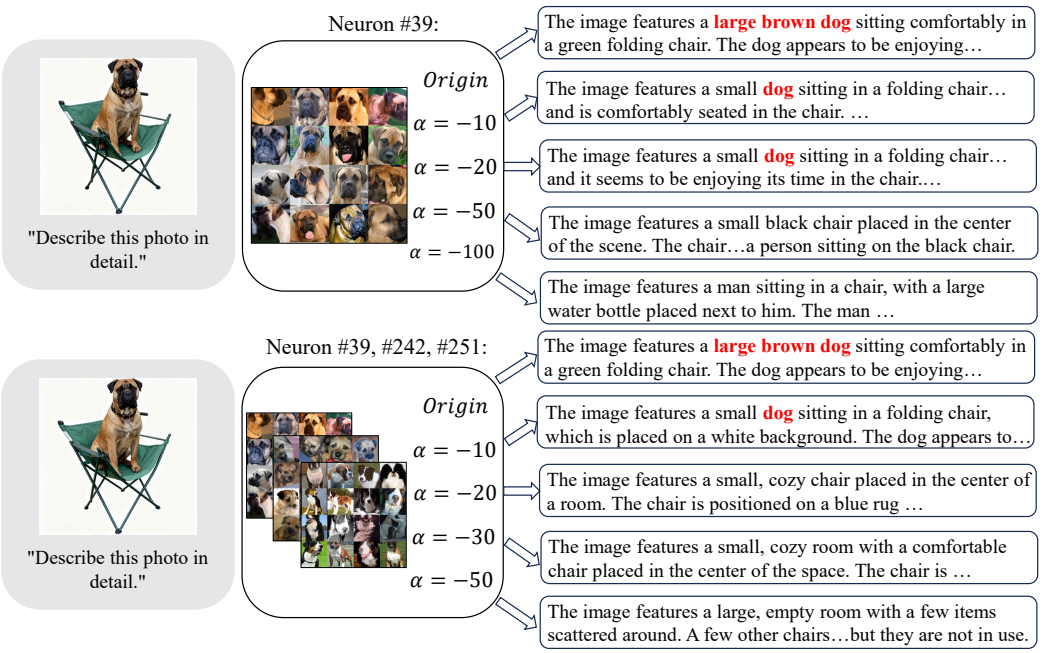


Figure 19: **Concept suppression in a complex context.** (a) Suppressing one dog-related feature requires a large negative intervention ($\alpha = -100$) before the concept disappears from the output. (b) Jointly suppressing three related features with $\alpha = -10$ each removes the concept with a smaller per-feature magnitude. This complements the insertion result by showing the value of coordinated updates for removing an unsupported concept.

F ALWAYS-ON NEURONS AND PERSISTENT ACTIVATIONS

Relation to persistent activation patterns. Several persistent-activation phenomena have been studied in neural networks, including *always-on neurons*, *register neurons* (Darcet et al., 2024; Jiang et al., 2025), *massive activations* (Sun et al., 2024), and *attention sinks* (Xiao et al., 2023; Kang et al., 2025). These phenomena share an intuitive property: some activations remain stable or large across many inputs and therefore cannot be explained solely by a local object in the current image. Our SAE analysis identifies a related empirical pattern in the visual representation. A small set of sparse features recurs across inputs, and in the analyzed representation these features often have activations around 10–80, whereas many of the top-40 responses are around 5–15. We use recurrence to define the always-on group for ANC, while using feature intervention to assess whether that group has a functional effect. This combination is important because persistence and causal relevance need not coincide: a feature can recur across images without being responsible for a particular hallucinated concept, while a less recurrent feature can be highly informative for the current image. We therefore use recurrence as an empirical property for compensation, without treating these features as universally harmful or functionally uniform.

Comparing unit definitions. Register neurons are defined by their architectural location, typically as particular MLP outputs within a transformer layer. Always-on neurons in our work are latent dimensions obtained by decomposing the visual representation with an SAE. They may summarize computation distributed across several architectural directions, and their identity depends on the SAE basis and the recurrence criterion. Thus, register-neuron analysis studies persistent architectural units, whereas our analysis studies recurring sparse visual factors. The two should not be treated as one-to-one counterparts.

Comparing recurrence criteria. The two groups also differ in how consistency is measured. Register-like activations may vary in token location or in the number of active positions, whereas our always-on group is a fixed set of SAE coordinates that frequently appears among the top responses across images. This is an operational grouping: its members may encode global statistics, recurring textures, or a stable activation background without sharing one computational role.

Comparing intervention interfaces. The intervention interfaces are also different. SAE features can be inspected through top-activating images and patches, manipulated before the visual tokens enter the multimodal connector, and evaluated through specific concepts. Register neurons are studied at their architectural locations and through their effects on transformer computation. The two views are complementary and could help determine whether a recurring SAE factor has a localized architectural source or summarizes information spread across layers and positions.

Role in CNS. Within CNS, the always-on group provides stable coordinates for activation-mass compensation. Contrastive steering changes the image-specific part of the code, and ANC redistributes the resulting change across these coordinates before the SAE code is decoded. Because their activations are generally large, the compensation can absorb the aggregate change with a small relative adjustment. Their role is therefore functional and stabilizing, rather than direct semantic suppression.

Future directions. More direct connections among always-on neurons, register neurons, massive activations, and attention sinks would require architectural tracing and causal experiments. Future work could compare these phenomena at matched layers, jointly measure their locations and norms, and test whether the same persistent factors remain stable across architectures, visual encoders, and multimodal connectors.