SAVE: Sparse Autoencoder-Driven Visual Information Enhancement for Mitigating Object Hallucination

Supplementary Material

A. Experimental Details

LURE Following LURE [43], we generate hallucinated objects for our visual information processing probe using GPT-3.5, which predicts objects likely to co-occur with the given image and prompt. Specifically, we prompt GPT-3.5 with: `"List three other objects that you think are most likely to appear with the objects in the scene described below."`

Model Architectures

- **LLaVA-1.6** For more details than those provided in 3.2, the SAEs are trained with a density factor of $\lambda = 5$, which is linearly increased from 0 during the first 5% of training steps to encourage sparsity. Training uses a total of 1.5B tokens with batches of 4096, shuffled to balance text and image tokens. The learning rate is 5e-5, decayed to zero over the final 20% of training, and optimization is performed using Adam.
- **LLaVA-NeXT** The Multimodal-SAE proposed by Zhang et al. [42] integrates a SAE into the 25th transformer layer of LLaVA-NeXT-LLaMA3-8B [17], where the hidden representation at that layer serves as the SAE input x. The SAE is trained on the LLaVA-NeXT supervised fine-tuning dataset [24], which consists of approximately 779,000 samples, using the AnyRes strategy for processing images of varying resolutions. mage and text inputs are preprocessed in the same way as during supervised fine-tuning.

  The SAE is configured with 2^{17} latent features and employs top-k sparsity, following Gao et al. [10] for the sparsity mechanism. We set $k = 256$ to match the activation patterns observed in Templeton et al. [36], promoting disentangled and semantically meaningful representations. Unless otherwise noted, all experiments are conducted using these SAE settings.
- **Qwen2-VL** Qwen2-VL-7B is trained following the exact same process as LLaVA-NeXT.

Configurations Table 1 reports results under the following configurations. For LLaVA-1.6, we evaluate POPE using steering at layer 24 with a strength (α) of 10, CHAIR at layer 24 with a strength of 15, and MMHal-Bench at layer 20 with a strength of 5. For LLaVA-NeXT, all benchmarks are evaluated with steering at layer 24, using a strength of 10 for POPE and MMHal-Bench and 15 for CHAIR. We set `max_new_tokens` to 2 for POPE, 256 for MMHal-Bench, and 512 for CHAIR.

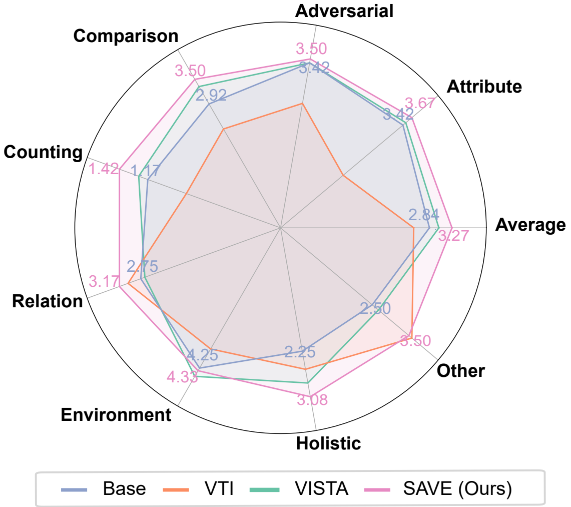


Figure S1. Task-wise MMHal-Bench performance on LLaVA-NeXT for different steering methods.

Steering We apply steering—i.e., modifying the encoded representation—only to the input query tokens, not to the tokens generated by the model.

The steering mechanism described in Section 4.3 applies to LLaVA-1.6. Since LLaVA-NeXT is trained differently, it requires a distinct steering strategy, which we also adopt for Qwen because it follows the same training process. In our experiments, this strategy is applied to both models, as formulated in the equation below, where T denotes tokens.

$$z(x) = \mathrm{ReLU}(W_{\mathrm{enc}}(x - b_{\mathrm{pre}}) + b_{\mathrm{enc}}) \tag{10}$$

$$\hat{z}(x)[T, j] = \alpha \quad \text{(steering)} \tag{11}$$

$$a(x) = \mathrm{TopK}(\hat{z}(x)) \tag{12}$$

$$x_{\mathrm{steered}} = W_{\mathrm{dec}} \cdot a(x) + b_{\mathrm{dec}} \tag{13}$$

Software and Hardware All experiments were conducted using PyTorch and an NVIDIA A40 GPU.

B. Detailed Results

MMHal-Bench MMHal-Bench reports scores across diverse tasks. As a detailed breakdown of Table 1, Figure S1 visualizes the task-wise performance on LLaVA-NeXT.

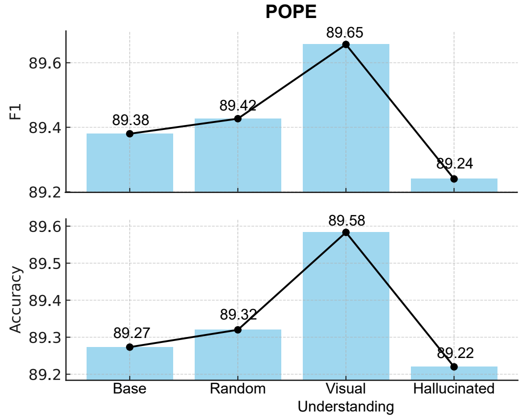


Figure S2. POPE evaluation of steering toward random, visual-understanding, and hallucinated features.

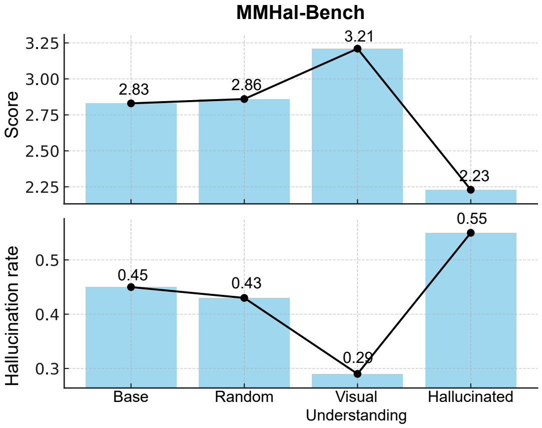


Figure S3. MMHal-Bench evaluation of steering toward random, visual-understanding, and hallucinated features.

Compared with the base model and various steering methods, SAVE consistently outperforms all baselines across every task.

Steering Model Behavior Additional results for the LLaVA-1.6 steering experiments presented in 4.3 are shown in Figure S2 and Figure S3. We report F1 and Accuracy for POPE, and GPT-4 Score with hallucination rate for MMHal-Bench. Across both benchmarks, steering along visual-understanding features mitigates hallucination, whereas steering along hallucinated features amplifies it. For this experiment, we follow the same configurations as in Table 1.

Layer-wise Steering Strength Ablation Table S1 and Table S2 provide the detailed results corresponding to Figure 6, for LLaVA-1.6 and Qwen2-VL, respectively. In Figure 6, we report the best performance per layer, selecting the steer strength that achieves the highest score. The optimal steer strength varies across layers: lower strengths (1.5 or 3) are more effective in early layers, while higher strengths (10 or 15) perform better in later layers. Through extensive ablations, we observe that overly strong steering at any layer can corrupt the model's output—e.g., generating repeated blanks or meaningless responses such as *"If you have any questions about the image, please provide more information"* (as observed in LLaVA-1.6, layer 8, steer strength 3). In the tables, a "–" indicates such corrupted outputs. Although the severity varies by layer, these findings highlight the importance of carefully selecting an appropriate steer strength to avoid response degradation.

Layer	Steer	CHAIR		POPE	
		CHAIR_S	CHAIR_I	F1	Acc
8	1.5	23.8	6.1	88.26	87.99
	3	–	–	–	–
12	3	21.6	6.0	89.27	88.90
	5	–	–	–	–
16	3	26.6	6.8	89.19	89.19
	5	–	–	89.15	88.78
20	5	31.6	7.9	89.31	89.21
	10	33.0	8.1	89.26	89.17
	15	29.0	10.3	89.06	88.97
24	5	32.2	8.4	89.41	89.21
	10	29.4	7.2	89.65	89.58
	15	21.4	5.4	89.55	89.46

Table S1. Performance of LLaVA-1.6 on CHAIR and POPE across layers and steering strengths.

Layer	Steer	CHAIR		POPE	
		CHAIR_S	CHAIR_I	F1	Acc
8	1.5	17.8	5.1	89.24	88.93
	3	20.2	5.9	89.22	88.89
12	3	23.0	6.5	88.48	88.92
	5	24.2	6.9	88.54	88.99
16	3	27.8	6.7	88.99	89.22
	5	26.8	6.5	89.01	89.24
20	5	23.8	6.4	88.75	89.03
	10	23.6	6.3	88.78	89.07
	15	24.4	6.8	88.79	89.08
24	5	20.8	8.0	85.16	86.61
	10	21.0	8.0	85.28	86.71
	15	21.4	8.2	85.28	86.71

Table S2. Performance of Qwen2-VL on CHAIR and POPE across layers and steering strengths.

Method	**Inference Time (s)**	**Generated Tokens**	**Total GFLOPs**	**FLOPs / Token**
Vanilla	7.998	215	38,166.48	177.52
VTI	12.553	248	38,662.83	155.90
VISTA	8.573	192	37,822.56	196.99
Devils	8.390	60	37,163.19	619.39
DeCo	10.850	220	67,947.49	308.85
SAVE (Ours)	8.863	174	40,281.54	231.50

Table S3. Inference efficiency comparison across methods.

Model	Method	Rec	OCR	Know	Gen	Spat	Math	Total
LLaVA-1.6	Base	45.2	36.5	33.6	36.0	35.5	26.5	42.2
	SAVE (Ours)	44.3	42.0	30.5	32.1	42.7	26.5	43.7
LLaVA-NeXT	Base	40.9	41.0	25.4	27.5	42.4	19.2	41.8
	SAVE (Ours)	43.1	43.1	30.2	32.5	43.1	22.7	42.2

Table S4. Per-task MMVet evaluation results on LLaVA-1.6 and LLaVA-NeXT.

Method	LLaVA-1.6	LLaVA-NeXT
Base	66.89	65.58
SAVE (Ours)	**70.04**	**66.81**

Table S5. A-OKVQA multiple-choice accuracy (%) on LLaVA-1.6 and LLaVA-NeXT.

C. Additional Results

Inference Efficiency We compare SAVE with several recent training-free methods—all reproduced on the LLaVA-NeXT backbone for consistency—using FLOPs and latency-based metrics. FLOPs are measured over the token generation process, which dominates overall inference computation. As shown in Table S3, SAVE achieves a favorable trade-off between efficiency and effectiveness: it generates a similar number of tokens while maintaining lower total FLOPs and FLOPs per token than strong baselines such as DeCo and Devils. Moreover, it runs faster than both DeCo and VTI, demonstrating efficiency in terms of both computation and latency.

VQA & General MLLM benchmark

- **MM-VET** MM-Vet [41] assesses visual understanding across six tasks—Recognition, OCR, Knowledge, Language Generation, Spatial Awareness, and Math. Table S4 shows that SAVE outperforms the base models on both LLaVA-1.6 and LLaVA-NeXT in terms of total score, using layer 20 with a steering strength of 3 for LLaVA-1.6, and layer 24 with a steering strength of 10 for LLaVA-NeXT.
- **A-OKVQA** A-OKVQA [33] is a crowdsourced benchmark of about 25K diverse questions that require commonsense reasoning about visual scenes beyond simple knowledge-base queries. As shown in Table S5, SAVE achieves higher multiple-choice accuracy than the base models, indicating that steering along visual understanding features not only mitigates hallucination but also enhances general visual understanding. Results are reported using LLaVA-1.6 with layer 20 and steering strength 5, and LLaVA-NeXT with layer 24 and steering strength 10.

Steer Strength We conduct an ablation study on steering strength (Figure S4). For LLaVA-NeXT at layer 24, we evaluate steer strengths of 3, 5, 10, 15, 20. The best results are obtained with a strength of 10 on POPE and MMHal-Bench, and 15 on CHAIR.

Extending the scope of SAE feature identification beyond object presence We conduct experiments on four visual reasoning tasks—existence (which corresponds to the object-presence signal used in the object-presence-only setting), count, position, and color—to evaluate the scalability of SAE-based feature extraction. In this setup, we use AMBER, a dataset organized by task type (see Figure S5), to obtain SAE activations, while evaluation is performed on MME to ensure generalizability. As shown in Table S6, LURE features naturally improve performance on the existence task, as they explicitly encode existence-related cues. This also explains why the existence scores of LURE and AMBER settings are identical. Interestingly, the LURE steering also improves performance on the color task, suggesting that visual cues associated with object existence implicitly contribute to color understanding as well.

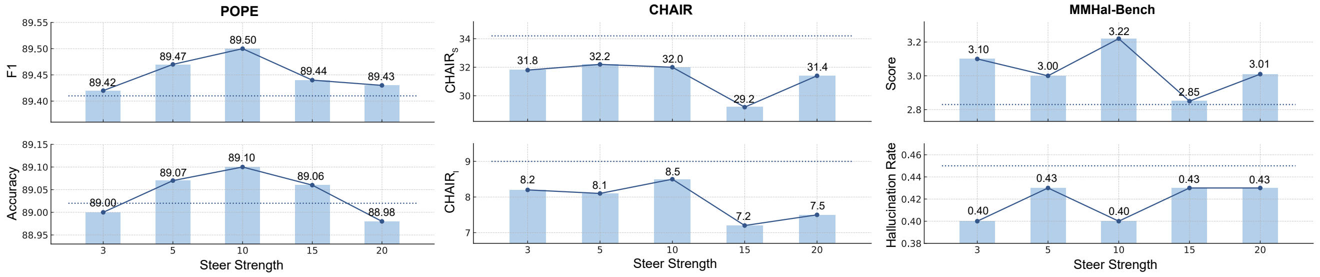
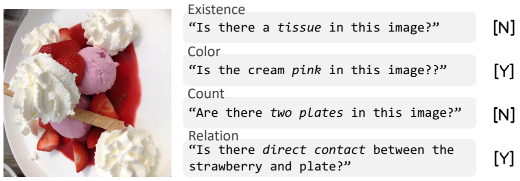


Figure S4. Steer strength ablation results for LLaVA-NeXT on POPE, CHAIR, and MMHal-Bench.

Task	Vanilla	LURE	Amber
Existence	195	200 (+5)	200 (+5)
Color	155	160 (+5)	160 (+5)
Count	115	115 (+0)	120 (+5)
Position	93.3	93.3 (+0)	98.3 (+5)

Table S6. Amber-derived SAE features outperform object-presence-only features particularly on **count** and **relation (position)** tasks.



Existence
"Is there a *tissue* in this image?" [N]

Color
"Is the cream *pink* in this image??" [Y]

Count
"Are there *two plates* in this image?" [N]

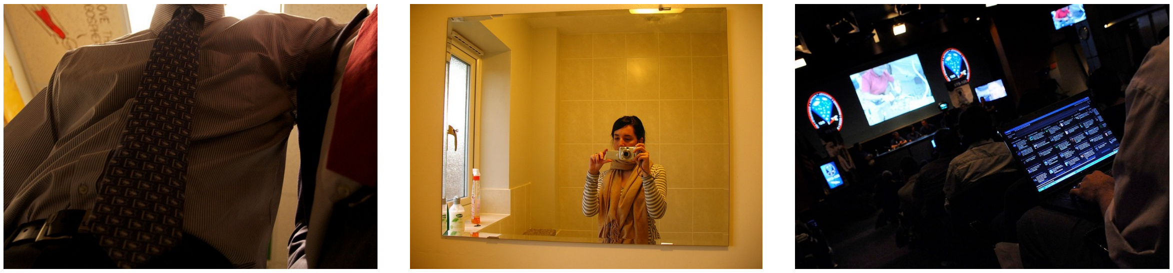
Relation
"Is there *direct contact* between the strawberry and plate?" [Y]

Figure S5. Examples of diverse question types from AMBER. For each task, both 'yes' and 'no' answer cases are included.

In contrast, SAE features extracted from the AMBER dataset—reflecting a richer variety of task-specific signals—lead to further gains in count and position. This suggests that when SAE directions encode broader aspects of the model's visual understanding, steering along those directions yields more comprehensive and balanced improvements across tasks. Overall, these findings demonstrate that our SAE-based latent steering approach is not limited to object presence, but can be effectively extended to support a wide range of visual reasoning objectives.

Qualitative Results for CHAIR Figure S6 shows qualitative CHAIR results. By steering along the identified visual-understanding features to enhance visual information, SAVE produces more visually grounded answers.

Statistical Test We conducted statistical significance testing on hallucination rates between LLaVA-1.6 Base (CHAIR$_S$=31.2, CHAIR$_I$=7.9) and LLaVA-1.6 SAVE (Ours) (CHAIR$_S$=21.4, CHAIR$_I$=5.4). Since the sample-level hallucination scores deviated from normality (Shapiro-Wilk $p < 0.001$), we employed the Wilcoxon signed-rank test, a non-parametric paired test assessing whether the median differences between models are systematically biased. At the instance level (CHAIR$_I$), SAVE achieved a modest yet statistically significant improvement ($W = 4998.0$, $p = 0.0188$, significant at $p < 0.05$). At the sentence level (CHAIR$_S$), the reduction was highly significant ($W = 2600.0$, $p = 0.00002$, significant at $p < 0.001$), demonstrating that SAVE consistently produces fewer hallucinated captions than the baseline, with the most pronounced effect at the sentence level.



LLaVA-Next: The person is seated in a chair, their body angled slightly to the left, creating a sense of depth in the image. Their right arm is casually resting on the armrest of the chair, while their left arm is draped over the back of the chair, adding a relaxed vibe to their otherwise formal appearance.

Ours: The background is blurred, but it appears to be an indoor setting with a wall that has some text on it, although the text is not clearly legible. The lighting in the image is soft and seems to be coming from the left side, casting a gentle shadow on the person's right side.

LLaVA-NeXT: The bathroom itself is bathed in a soft, warm light that accentuates the beige tiles and the white sink. The sink, located on the right side of the image, is adorned with a few toiletries - a bottle of soap and a toothbrush holder.

Ours: The bathroom has a beige color scheme with white tiles on the walls and a window with frosted glass to the left. On the window sill, there are two bottles, one of which appears to be a bottle of hand soap.

LLaVA-NeXT:The room itself is well-lit, with multiple screens and a clock visible in the background. The presence of these screens and the clock suggests that this is a professional setting, possibly a conference hall or seminar room.

Ours: The setting seems to be a conference room or seminar hall, as there are multiple people seated in the background. The room is well-lit, with several screens and monitors placed around the room, likely for presentations or other multimedia content.

Figure S6. Qualitative comparison between vanilla LLaVA-NeXT and our method on the CHAIR benchmark. Our steered model generates visually grounded captions, while vanilla LLaVA-NeXT exhibits object hallucination. Hallucinated words are highlighted in red, and correct responses are shown in green.