Locate-then-Sparsify: Attribution Guided Sparse Strategy for Visual Hallucination Mitigation

Supplementary Material

6. Details of the construction of the dataset

In this section, we introduce the details of how to construct the Bi-granularity Dataset.

At first, to preserve generalization, the data used for dataset construction and the data used for experiments are strictly disjoint. Particularly, for data selected based on CHAIR and POPE, we use data from the train split of MSCOCO. For data selected from Antidote, we do not use these data for evaluation.

Secondly, we explain how to get a single piece of data. As an example from CHAIR, the data instance is generated by an LVLM. We use LLaVA-v1.5-7B to produce a response according to the CHAIR benchmark, as illustrated in Fig. 6. We then apply CHAIR's evaluation criteria to detect hallucination and annotate the instance under our two-level scheme (token- and sentence-level). And then a piece of data is generated. If the responses don't have hallucination, they are just not selected.

Finally, to balance two levels of data, we select 100 sentence-level samples and 100 token-level samples. All data are manually inspected to ensure accuracy.

7. Implementation details of LTS-FS

Hyper-parameters. The strength control parameters of s^l_{tok}: $\lambda_{cue}, \lambda_{pos}, \lambda_{hall}$ is set to be 1. The mask threshold r_s is selected to be 0.5, as shown in Tab. 5 of the main text.

Environment. All the experiments are conducted on one A100 80G. For the 7B model, two RTX3090 24G can replace an A100. For detailed Python requirements, please refer to our released code.

8. Implementation Settings of CHAIR Results

Generation Setting. Here we set the generation config as follows: **Max_New_Tokens=128**, **num_beams=1**, and **sampling=False**.

Compared methods. We employ the default parameters and settings as reported in the original papers.

9. Generation Capability.

To evaluate general capability more comprehensively, we perform an evaluation using a broader benchmark called CLAIR [6]. This result in Tab. 7 shows that LTS-FS achieves a better trade-off between hallucination mitigation and general capability preservation.

Table 7. Trade-off between hallucination mitigation and general capability preservation.

Method	CHAIRs	POPE acc	details	CLAIR
Original	53.0	77.63	5.23	80.03
nullu	50.2	79.11	5.51	75.00
LTS-FS(nullu)	46.8	79.92	6.23	82.74
Soft Gating	46.7	79.5	6.26	83.64

10. More details of POPE results

Generation Setting. Here we set the generation config as follows: **Max_New_Tokens=16**, **num_beams=1**, and **sampling=False**.

Compared methods. We employ the default parameters and settings as reported in the original papers.

Total Results. The total results is shown in Tab. 13. Across all settings, our LTS-FS framework achieves the best accuracy and F1, demonstrating consistent effectiveness in hallucination mitigation. Compared with the original feature-steering methods, applying LTS-FS consistently improves both VTI and Nullu on hallucination-related metrics. Although LTS-FS and the other methods trade wins on recall, LTS-FS consistently maintains higher precision. Since, in hallucination evaluation, precision is more indicative of mitigation quality, this further supports the strong performance of our approach.

11. More details of MME results

We report the MME numerical results in Tab. 8. The numerical results demonstrate that LTS-FS can strongly increase the mitigation ability of feature steering methods. Specifically, across the subset most related to hallucination: Count, and Position, LTS-FS achieves great improvements, highlighting its effectiveness in enhancing feature-steering–based mitigation.

MME includes not only perception-related tasks but also recognition-related tasks. We report these results in Tab. 9. Despite the sparsity selection emphasizes hallucination related cues rather than recognition factors, LTS-FS still produces improvements on recognition-related tasks.

12. Time Analysis

There are two time cost analyses: the time to apply methods and the time for inference. The time to apply methods is the time to employ a hallucination mitigation method in a specific LVLM. As an example, in order to apply VTI to LVLMs, the direction vector needs to be computed, and

Table 8. Results on all MME perception-related tasks.

Method	*Existence*	*Count*	*Position*	*Color*	*Posters*	*Celebrity*	*Scene*	*Landmark*	*Artwork*	*OCR*	**Total**
Regular	182	118	105	151	118	112	145	131	108	78	1248
Nullu	190	122	106	157	128	118	148	130	114	121	1334
LTS-FS(Nullu)	195	153	128	157	130	127	155	131	113	123	1412

Table 9. Results on all MME recognition-related tasks.

Model	Method	Common Sense Reasoning	Numerical Calculation	Text Translation	Code Reasoning	Total
LLaVA-1.5-7B	Regular	110	50	50	71	281
	Nullu	113	59	75	77	324
	LTS-FS + Nullu	120	59	75	80	334

Table 10. Time analysis comparison of different hallucination mitigation strategies. VCD represents a decoding-based method. Nullu represents a feature-steering-based method.

Method	**Preparation Cost**	**Inference Cost**
Regular	–	1.31s
VCD	0s	3.14s
Nullu	30mins	1.37s
Ours	90mins	1.34s

Table 11. Ablation study of indicators. HI, CI, PI respectively indicate hallucination indicator, cue indicator, and position indicator.

Settings	C_S	C_S	Recall	Length
Regular	53.0	13.9	77.2	98.0
w/o HI	52.0	14.0	76.9	97.4
w/o CI	48.2	13.6	77.1	95.7
w/o PI	47.6	13.7	76.9	94.3
LTS-FS(Nullu)	**46.8**	**13.5**	76.6	93.2

Table 12. Results of the generalization test of our framework. We use LLaVA-v1.5-7B to conduct this experiment. C_S and C_I is the $CHAIR_S$ and $CHAIR_I$ under CHAIR benchmark. ACC and F1 mean the accuracy and F1 score in the GQA subset on POPE.

Settings	C_S	C_I	Acc	F1
Regular	53.0	13.9	75.47	79.83
MSCOCO→GQA	—	—	77.31	79.57
GQA→MSCOCO	49.5	13.2	—	—
Antidote→GQA	—	—	77.28	80.12
Antidote→MSCOCO	49.8	13.7	—	—
LTS-FS(Nullu)	46.8	13.5	77.15	80.63

the layer should be adjusted. This whole time is the time to apply methods. For our method, the time to apply the methods contains two parts. First, we need layer-wise attribution to select specific layers. Second, we need to apply feature steering methods based on these sparse layers. The second part time is almost the same as the original feature steering methods, which can be completed in under 30 minutes. The first part is the attribution process, which is time-consuming. For LLaVA-v1.5-7B, it takes about 1–2 hours on a single A100 80 GB GPU.

As for the time for inference, our framework is based on feature steering methods. Therefore, the time for inference is comparable with regular generation. Comparison is shown in Tab. 10. Despite requiring a longer preparation phase, the additional cost is reasonable, as it avoids the extra inference time latency that would otherwise accumulate during decoding and further highlights the inherent advantages of feature steering techniques.

13. Ablation Study about Indicators

In this section, we discuss the effect of the three indicators on sentence-level hallucination attribution. The result is shown in Tab. 11. We investigate the effect of removing each indicator in turn and find that *w/o cue indicator* and *w/o position indicator* yield only small changes, whereas *w/o hallucination* causes a much larger decline, indicating that hallucination token attribution is paramount, with cue and position still providing auxiliary gains.

14. Discussion about Generalization

To assess generalization beyond the construction sources, we evaluate on datasets whose distributions differ from those used to build our bi-granularity labels. Although the construction leverages CHAIR, POPE, and Antidote, we additionally report results on MME and LLaVA-Bench, which serve as an out-of-distribution dataset of overall capability. We also run a decoupled calibration evaluation protocol: layer scores and weights are calibrated on one source (e.g., CHAIR on MSCOCO), then frozen

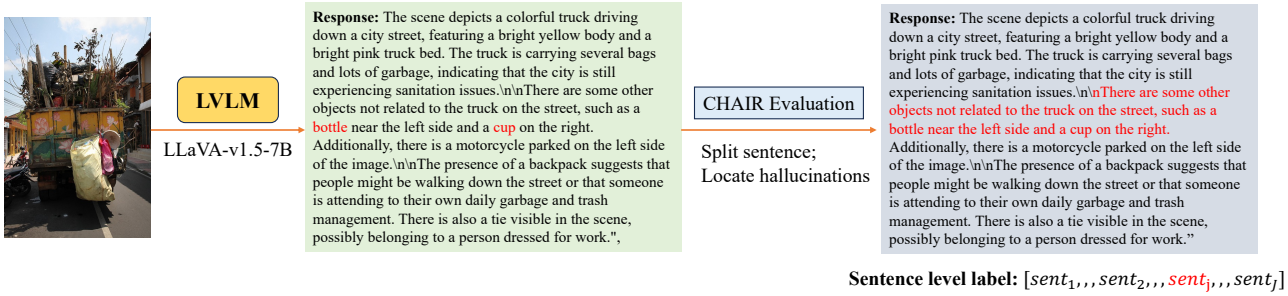

Figure 6. A sample generation based on CHAIR benchmark

Figure 7. More examples on LLaVA-Bench.

and applied to a different target set for evaluation (e.g., POPE-GQA or Antidote). Concretely, CHAIR relies on MSCOCO; POPE uses MSCOCO and GQA; Antidote uses its own corpus. We therefore test cross-dataset pairs such as MSCOCO→GQA to verify transfer. The results are shown in Tab. 12). MSCOCO→GQA denotes calibrating attribution on MSCOCO and evaluating on the POPE–GQA subset, and GQA→MSCOCO means attribution based on the GQA dataset and evaluation on the MSCOCO dataset under the CHAIR benchmark. Despite calibrating on only part of the data, our framework typically delivers additional gains. The findings suggest that our improvements are driven by intrinsic generalization capacity, not by overfitting to a particular data distribution.

15. More cases in LLaVA-bench

More case studies on the LLaVA-bench are presented in Fig. 7, which demonstrates the effectiveness of our framework in hallucination mitigation. In particular, color and count attributes are given greater emphasis, thereby avoiding hallucinations in these aspects.

16. GPT4v-Evaluation prompt

Following VCD, the prompt for GPT4v-aided evaluation is shown in Fig. 8. The GPT4v receives three types of LVLM's responses and then generates output. Then we collect the output from GPT4v and finally report the average accuracy and detailedness.

17. Limitation and future work

Although our approach can be effectively ported to feature-steering methods and achieves strong hallucination mitigation, there is still room for development. Since existing feature steering techniques have not been evaluated on larger 70B-scale models, extending our method to 70B models remains a challenge. We aim to extend our framework to larger models and further investigate its impact across additional multimodal domains.

Table 13. Average POPE results with Random and Popular.

Setting	Model	Method	Accuracy ↑	Precision↑	Recall ↑	F1 Score ↑
Random	LLaVA-v1.5-7B	Regular	85.37	80.77	93.22	86.47
		VCD	86.55	84.02	90.69	87.16
		AGLA	85.32	83.56	91.34	86.77
		Nullu	86.35	84.36	91.09	86.28
		VTI	84.84	80.02	93.36	86.08
		LTS-FS(Nullu)	**87.13**	84.69	91.02	**87.64**
		LTS-FS(VTI)	86.77	84.13	91.00	87.32
	LLaVA-v1.5-13B	Regular	81.91	75.84	93.82	83.85
		VCD	82.27	75.97	92.68	83.76
		AGLA	82.64	76.19	93.16	83.58
		Nullu	83.24	77.93	92.89	84.73
		VTI	84.08	76.29	93.04	83.82
		LTS-FS(Nullu)	83.96	78.89	93.85	85.56
		LTS-FS(VTI)	**86.59**	82.35	93.47	**87.48**
	Qwen-VL2.5-7B	Regular	85.32	96.38	73.57	84.03
		VCD	85.94	97.13	74.11	83.89
		AGLA	86.02	96.56	73.65	83.63
		Nullu	85.82	97.17	73.93	83.73
		VTI	85.49	96.85	73.51	83.37
		LTS-FS(Nullu)	**86.21**	97.09	74.78	**84.31**
		LTS-FS(VTI)	86.04	97.23	73.64	83.87
Popular	LLaVA-v1.5-7B	Regular	77.52	71.45	93.22	80.71
		VCD	79.09	73.21	92.17	81.23
		AGLA	78.67	75.39	89.02	81.47
		Nullu	79.42	74.45	91.04	81.67
		VTI	77.03	70.90	93.36	80.40
		LTS-FS(Nullu)	**80.09**	75.28	91.07	82.20
		LTS-FS(VTI)	79.96	75.25	91.14	**82.25**
	LLaVA-v1.5-13B	Regular	78.40	71.78	93.76	81.30
		VCD	79.38	72.24	92.47	82.01
		AGLA	80.11	72.88	92.16	82.32
		Nullu	80.88	74.91	93.02	82.97
		VTI	79.22	73.26	93.04	81.83
		LTS-FS(Nullu)	81.46	75.62	92.93	83.42
		LTS-FS(VTI)	**81.77**	75.58	93.47	**83.58**
	Qwen-VL2.5-7B	Regular	83.31	91.14	74.58	81.68
		VCD	83.19	90.27	74.18	81.95
		AGLA	83.34	90.69	74.53	81.86
		Nullu	83.06	91.20	74.04	81.27
		VTI	82.74	90.86	73.51	80.88
		LTS-FS(Nullu)	**83.59**	91.12	76.18	**82.55**
		LTS-FS(VTI)	83.35	90.96	73.64	81.38
Adversarial	LLaVA-v1.5-7B	Regular	70.13	64.14	93.22	75.85
		VCD	71.48	66.28	89.62	76.04
		AGLA	71.63	66.59	90.13	75.89
		Nullu	71.57	66.06	90.53	76.17
		VTI	69.40	63.46	93.36	75.42
		LTS-FS(Nullu)	72.62	65.99	90.62	76.22
		LTS-FS(VTI)	**73.04**	67.37	91.24	**77.32**
	LLaVA-v1.5-13B	Regular	71.07	64.60	93.83	76.47
		VCD	71.73	63.61	94.23	75.76
		AGLA	72.27	64.14	93.56	75.48
		Nullu	72.43	66.08	92.44	77.04
		VTI	71.77	65.58	93.01	76.80
		LTS-FS(Nullu)	73.06	67.01	92.96	78.36
		LTS-FS(VTI)	**73.78**	67.51	93.47	**79.91**
	Qwen-VL2.5-7B	Regular	80.17	85.21	73.64	78.93
		VCD	80.56	85.31	75.07	79.51
		AGLA	80.92	85.73	74.72	79.14
		Nullu	80.74	86.32	74.24	79.32
		VTI	80.19	85.75	73.51	78.70
		LTS-FS(Nullu)	**81.11**	86.14	75.07	**79.83**
		LTS-FS(VTI)	80.92	85.94	73.64	79.46

Description:
AI that scores image description accuracy and detailedness.

Instructions:
You are an AI designed to evaluate and score the performance of three AI assistants in describing a given image. Your primary focus is on the accuracy and detailedness of their descriptions. You will assess the accuracy by checking for hallucinations - any part of the description that is inconsistent with the image content. For detailedness, you will consider how rich the response is in necessary details, excluding any hallucinated parts. You will provide scores on a scale from 1 to 10 for each assistant separately, based on these criteria. After scoring, you will offer an explanation for your evaluation, ensuring it is free from bias and not influenced by the order of presentation of the responses.

Input format:

[Assistant 1]
{Response 1}
[End of Assistant 1]

[Assistant 2]
{Response 2}
[End of Assistant 2]

[Assistant 3]
{Response 3}
[End of Assistant 3]

Output format:

Accuracy:
Scores of the three answers:
Reason:

Detailedness:
Scores of the three answers:
Reason:

Figure 8. Prompt of GPT-4V Evaluation.